

AD _____

Award Number: DAMD17-98-1-8649

TITLE: Use of Novel Technologies to Identify and Investigate
Molecular Markers for Ovarian Cancer Screening and Prevention

PRINCIPAL INVESTIGATOR: Nicole D. Urban, Sc.D.

CONTRACTING ORGANIZATION: Fred Hutchinson Cancer Research Center
Seattle, Washington 98109-1024

REPORT DATE: October 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20021001 077

REPORT DOCUMENTATION PAGEForm Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE October 2001	3. REPORT TYPE AND DATES COVERED Final (1 Oct 98 - 30 Sep 01)	
4. TITLE AND SUBTITLE Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention			5. FUNDING NUMBERS DAMD17-98-1-8649	
6. AUTHOR(S) Nicole D. Urban, Sc.D.				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Fred Hutchinson Cancer Research Center Seattle, Washington 98109-1024 E-Mail: nurban@fhcrc.org			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES Report contains color				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited				12b. DISTRIBUTION CODE
13. ABSTRACT (Maximum 200 Words) The purpose of this study is to identify novel genes that encode proteins that can be used to detect ovarian cancer before it spreads outside the ovary and becomes incurable. The goal is to assemble a panel of known and novel ovarian tumor markers that may form the basis of a cost-effective, serologic screening test for early stage ovarian tumors. The research encompasses the use of two novel technologies to identify such genes. In Project 1 we use HDAH to identify genes that are over-expressed in ovarian cancer tissue. In Project 2 novel ovarian tumor antigens are being identified by SEREX. We have identified a large number of genes that are over-expressed in ovarian cancer tissue relative to the ovarian tissue obtained from women without cancer or ovarian pathology. We have also identified several oncogenic proteins that elicit antibodies detectable in the blood of some ovarian cancer patients. These discoveries provided the foundation for ongoing work in early detection of ovarian cancer, funded by the NCI as part of a SPORE in ovarian cancer.				
14. SUBJECT TERMS Ovarian Cancer Screening, Gene Expression, Serum Antibody, High Density Array Hybridization (HDAH), SEREX				15. NUMBER OF PAGES 309
				16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Final Report for Grant DAMD17-98-1-8649

October 1, 1998 – October 31, 2001

Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian
Cancer Screening and Prevention

Nicole Urban, ScD
Principal Investigator

TABLE OF CONTENTS

Front Cover...	1
SF 298 Report Documentation Page	2
Table of Contents	3
Report Cover	4
Introduction	5
Body	9
Key Research Accomplishments.....	71
Reportable Outcomes	73
Conclusions	79
Appendix A	80
Appendix B	86
Appendix C	88
Appendix D	98
Appendix E	137
Appendix F.....	141
Appendix G	152
Appendix H.....	235
Appendix I.....	262
Appendix J	264
Appendix K.....	268
Appendix L	271
Appendix M	273
Appendix N.....	293
Appendix O	295

Final Report for Grant DAMD17-98-1-8649

October 1, 1998 – October 31, 2001

**Use of Novel Technologies to Identify and Investigate Molecular
Markers for Ovarian Cancer Screening and Prevention**

**Nicole Urban, ScD
Principal Investigator**

INTRODUCTION

Three years ago, we were awarded funding for a study entitled “Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian Cancer Screening and Prevention” (DAMD 17-98-1-8649), to conduct a systematic search for novel genes and gene products associated with ovarian cancer. In this report we summarize our progress. In the original statement of work, 5 major tasks were identified for Project 1, 5 major tasks were identified for Project 2, and 20 major tasks were identified for the Statistical, Clinical and Laboratory Coordinating Core. These tasks are listed in a table included in Appendix A. Included as Appendix B is a timeline detailing project progress during the final no-cost extension year. Products in the form of published or submitted manuscripts are provided in the Appendices.

Our goal is to improve ovarian cancer outcomes, including morbidity and identification of novel genes associated with ovarian cancer. This interdisciplinary effort, which required three years of work including a one-year no-cost extension (10/98-9/01), addresses gene discovery as it relates to risk-assessment and early detection. Our purpose is to identify novel genes that encode proteins that can potentially be used to detect ovarian cancer before it spreads outside the ovary and becomes incurable. The goal is to assemble a panel of known and novel ovarian tumor markers that may form the basis of a cost-effective, serologic screening test for early stage ovarian tumors. Our DOD-funded work has led directly to funding by the NCI of a SPORE in ovarian cancer. The scope of the research encompasses the use of two novel technologies to identify such genes. It included two research projects, described below.

HDAH. In Project 1, entitled “Characterization of Genes Overexpressed in Malignant Ovarian Neoplasia by High Density Array Hybridization” (DAMD 17-98-1-8649), we used high density array hybridization (HDAH) to identify genes that are over-expressed in ovarian cancer tissue. Drs. Leroy Hood, Nancy Kiviat and Michel Schummer used high-density cDNA array hybridization (HDAH) to compare the expression of genes in normal and in neoplastic ovarian tissue. Genes that are highly expressed in malignant tissue, but expressed at low levels in benign and normal tissue, are potential candidates for development as diagnostic markers. We have built our own libraries from unique ovarian tissues for the hybridization work to ensure that we will discover *novel* genes. We have found many over-expressed genes using HDAH, from which the most promising have been selected for further work-up.

For example, a top candidate for development was HE4, an epididymal gene that maps to a region of the genome that is a hot spot for changes in ovarian and other cancers. This region is found to be amplified in ovarian and breast cancer, as well in some glioblastomas. It is possible that this amplification of 20q12-13.1 in ovarian cancers causes HE4 to be over-expressed. Dr. Schummer has collaborated with Drs. Ingegerd and Karl-Erik Hellstrom to develop an assay (sandwich ELISA) for HE4. Another is a mesothelin-related gene, a 40-kDa glycoprotein present on the surface of many different malignancies including the majority of mesotheliomas and ovarian cancers. Drs. Ingegerd and Karl-Erik Hellstrom have recently developed an assay to detect mesothelin in serum. These assays are now being evaluated for their contribution to a panel of markers to detect ovarian cancer as SPORE grant developmental studies.

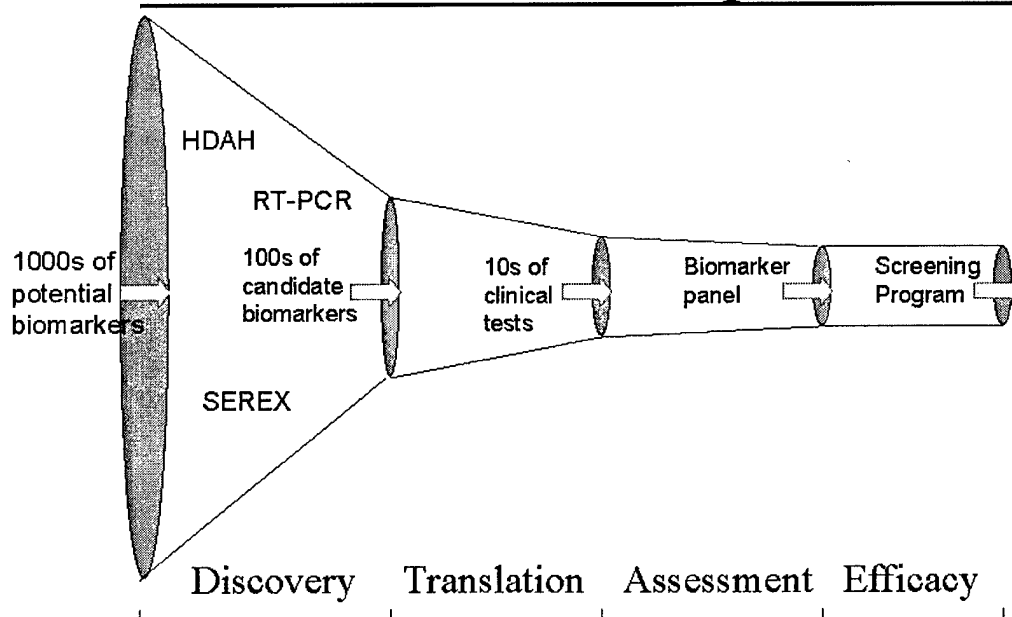
SEREX. In addition to HDAH, we have used SEREX, a novel serological method that identifies immunogenic gene products for which antibodies are present in the sera of women with ovarian cancer but not in those of controls. In the project entitled “Antibody Immunity to Cancer Related Proteins as a Serologic Marker for Ovarian Cancer” (DAMD 17-98-1-8649) Drs. Brad Nelson and Mary L. Disis are using SEREX to identify antibodies to novel cancer-associated proteins. This technology involved (1) construction of a bacterial cDNA expression library from a pooled tissue sample representing the tumors of selected ovarian cancer patients, (2) probing of the library by immunoblot with serum from both cancer patients and control individuals, and (3) identification of bacterial colonies that are recognized by serum antibodies from cancer patients but not normals. Antibodies found only in the sera of cancer patients are candidates for development, validation, and evaluation for inclusion in a set of markers to be used as a first-line screen in the early detection of ovarian cancer. In addition, this technology defines genes by the immunogenic proteins that they encode. These genes would be possible candidates for DNA vaccines. We have identified several oncogenic proteins using SEREX and are evaluating the most promising for use as markers of ovarian cancer risk.

Core. Novel genes identified by HDAH have been evaluated for development as tumor markers. Similarly, antigens identified by SEREX are being further evaluated using purified proteins and larger numbers of normal, benign and cancer serum in an ELISA format. Statistical analyses have been used to determine the sensitivity and specificity of HDAH- and SEREX-defined tumor markers for the detection of early-stage ovarian cancer. Towards this goal, serum antibody responses measured by ELISA to the known tumor antigens p53 and HER2 have been evaluated in patients with ovarian cancer versus normal controls.

Research Pipeline. Separate tasks were identified for each project and core as being imperative to the successful completion of this grant. The study was developed as the first phase (Discovery) of a four part research pipeline (Translation, Assessment and Efficacy as the second, third and fourth phases). Four of the 5 projects in our SPORE continues work that began with DOD funding, in the hope of realizing our goal of improving ovarian cancer outcomes. SPORE Projects 1 and 2 are designed to provide information that will enable us eventually to reduce cancer mortality and incidence through treatment and prevention interventions respectively; Project 1 uses the HDAH technology to identify genes associated with resistance to chemotherapy, and Project 2 seeks to develop vaccines targeting Her2/neu and EGFR. Eventually we hope to develop vaccines against targets identified by our HDAH and SEREX technologies. SPORE Project 3 provides the methods that will be needed to conduct screening and prevention trials when appropriate screening and prevention interventions have been identified. SPORE Project 4 provides the intervention for use in a screening trial. SPORE Projects 3 and 4 make use of the markers that we have found, in combination with previously known markers, in panels designed to measure risk of ovarian cancer.

The process of marker discovery, validation and application is depicted below.

A Research Pipeline



Accomplishments

Gene Discovery

We have identified several potential ovarian cancer marker genes, some of them previously known to be related to cancer (WFDC2, FOLR1, CD24, KRT8, MUC1, LCN2, S100A11, S100A6, IFI27, ERBB2), some with no previous role in cancer (ELF3, GPR39, SLPI, MSLN), some matching to ESTs (22) and some being novel genes (8). These genes and their proteins have continued to be evaluated by research laboratories in our ovarian SPORE grant, with NCI supplemental and SPORE developmental funds. The antibody generation is being carried out at the Pacific Northwest Research Institute in Seattle by I. and KE Hellström); in situ hybridization on tissue sections and transcript quantitation on cells from peritoneal washes are conducted in the laboratory of co-Investigator N. Kiviat at the University of Washington. Markers from this study are also being employed in a study funded through the Ovarian SPORE where multiple markers, including CA125, will be evaluated together (SPORE Project 4 by Martin McIntosh).

Protein Discovery

In addition to the known tumor antigens p53 and HER2/neu that were evaluated, a large number of novel candidate antigens that are immunogenic in ovarian cancer have been identified by SEREX. We are now poised to evaluate serum antibody responses to the combined panel of antigens using large numbers of sera from patients with malignant and benign ovarian disease and normal controls. In follow up to their SEREX work, Nora Disis and Brad Nelson are exploring the translational potential of about 15 oncogenic proteins for which antibodies are present in the serum of ovarian cancer patients but not in the serum of healthy controls. In collaboration with Stan Riddell of the FHCRC they are developing methods to realize the potential of translational opportunities in immunology using these proteins, beginning with Her2/neu and EGFR. This new work is funded by a SPORE developmental research project.

Evaluation of Novel Technologies Data

It is imperative in the course of gene and marker discovery that the statistical properties of potential markers be well characterized prior to adoption for further development. With data generated by this study, Garnet Anderson, in collaboration with Margaret Pepe at FHCRC, have developed statistical methods for analyses of gene expression data, specifically, methods to rank genes or proteins with regards to differential expression between tissues. Their methods place confidence intervals around the rankings and allow estimation of sample sizes required to rank genes.

Publications

Knowledge gained from this grant has been shared with the scientific community. A total of 11 manuscripts were developed during the funding period, with 5 manuscripts utilizing data generated from this study accepted for publication during the no-cost extension. These papers are cited in the reportable outcomes section and are included as appendices.

Interdisciplinary interaction is the key to progress.

Clinical, laboratory and public health scientists were involved in every component of this grant. Interdisciplinary collaboration has been highly productive, particularly between the statisticians and laboratory scientists, and this model has been carried forward in our SPORE grant. Interdisciplinary interaction occurs across all components of the SPORE, with investigators organized into interdisciplinary teams that including discovery, translation, early detection, prevention, treatment, outcomes, and minority affairs. Each team includes laboratory scientists, clinicians, statisticians, and representatives of other relevant disciplines.

Project 1

Identification of Potential Markers for Population Based Screening in Ovarian Cancer: Characterization of Differential Gene Expression in Malignant Neoplasia by Use of High Density Array Hybridization

Nicole Urban, ScD, Leroy Hood, MD, Ph.D.; Michel Schummer, Ph.D.

INTRODUCTION

It is well established that the set of genes expressed in tumor cells differ from that expressed their normal counterparts in both a qualitative (different genes expressed) and quantitative fashion. These differences in gene expression, and specifically overexpression are exceedingly common in cancers at the level of mRNA and provide a logical basis for cancer screening assays. We have proposed a rapid and accurate approach to identification of genes which are overexpressed in ovarian cancers and which are likely to be of interest for use in ovarian cancer screening assays. We have used multiple rounds of cDNA array hybridization to identify a subset of a few dozen genes which are overexpressed in a high percentage of early and late stage ovarian cancers but not in normal tissues. Once such genes were identified by array hybridization, they were sequenced and by comparing sequences to described sequences on public databases, we were able to target those which appear to code for secreted and/or for transmembrane proteins for further characterization by quantitative RealTime PCR on tissues and circulating cells from peritoneal washes. For two of these, WFDC2 and MSLN, ELISAs were performed to establish the presence of their proteins in sera of ovarian cancer patients. When used in combination with CA125, both proteins add specificity and sensitivity to CA125.

BODY:

Work proposed:

Task 1. Generation of representative cDNA arrays:

- Three cDNA libraries will be generated from normal, metastatic and late stage neoplastic ovarian tissues.
- These libraries will then be used to construct first generation solid phase membrane arrays containing 100,000 clones.

Task 2. Primary Characterization of Normal and Neoplastic Ovarian Tissue:

- Hybridization of the first generation membranes with cDNA probes derived from 12 normal (pre and post menopausal ovarian tissue, 3 peripheral blood samples, peripheral blood cell culture and 1 liver tissue, 4 2 benign cystadenomas, 1 early stage and 12 late stage ovarian serous adenocarcinomas
- Evaluation of hybridization results and selection of 2,000-3,000 genes overexpressed in malignant tissues.
- These clones will be used to construct second generation cDNA arrays.

Task 3. Further Characterization of Gene Expression in Normal and Neoplastic Ovarian Tissue:

- Hybridization of the second generation arrays with cDNA from tissues used in Task 2, plus 29 additional normal tissues (20 ovarian and 9 skeletal muscle controls), 15 cystadenomas, 20 additional early and 20 late stage ovarian serous adenocarcinomas.
- Evaluation of hybridization results and selection of ~400 genes that show a high degree of overexpression in at least 75% of tumors examined.

Task 4. Characterization of highly expressed genes associated with cancer:

- Sequence determination of the ~400 overexpressed ovarian cancer-associated genes identified in Task 3.
- Confirmation of tissue specific expression using RT-PCR and Northern blot techniques.
- Selection of clones with overexpression in ovarian cancers negative for overexpression of p53, Her2/neu and c-myc transcripts for further analysis by serum-based detection technologies in Project 2.

Task 5. Final analyses and report writing:

- Final analyses of serum-based patient screening assays will be performed.
- A final report and initial manuscripts will be prepared

Table 1 - Work Flow

	Task 1	Task 2	Task 3 ...	Task 4 ...	Task 5
total tissues / sera	32		64		16
normal non-ovarian	1		1		
PBL	3				
PBL culture	1		1		
ovarian fibroblast cultures					
pool of fetal ovaries					
OSE					
normal ovaries	11		23		8
omentum or fallopian tubes	1		1		
benign tumors	2		7		
borderline tumors					
stage I mucinous	1		1		
stage III serous	8		24		8
stage IV serous	4		5		
metastatic tissues					
blinded ovary			1		
ovarian cell lines					
breast cell lines					
cervical cell lines					
endometrium cell line					
97,000	883	1390	114	78	5
≈ 32,000 genes	45 Novel 366 ESTs 467 Known	139 Novel 560 ESTs	8 Novel 30 ESTs	1 Novel 22 ESTs	1 Novel 1 EST
			AKT1	1-4	1 EST
			AKT2	actin beta	CD24
			c-jun	bamacon	ESE-1
			Calvasculin	BRCA1	Calgizzarin
			Cyclin C	BRCA2	ESO-1
			EDN1	c-myc	GPR39
			ESR1	CCR2	HE4
			ESR2	E16	Folate BP
			HGF	Ferritin H	Her2/neu
			HSD3B2	GA733-1	Keratin8
			IL-8	GAB2	Lipocalin2
			Ku70	GAPDH	Mesothelin
			Lot1	IGF BP2	Mucin1
			MAGE-4	IGF2	p27
			MET	Kadereit	PAX2
			MIS	Ku80	SLPI
			p73	MAT1	
			PIK3CA	MCAF	PLTP
			StAR	MDC15	progBP
			STK11	Nup88	PTEN
			TADG-14	oviduct-gp	RIG-E
			UNC119	p53	Ryudocan
					SAS
					ST5

Starting from an array of 100,000 clones hybridized with probes from 32 tissues on the left, we selected 883 genes to be potential marker genes and combined them with 507 genes compiled from previous membrane arrays and other sources, to form a glass array (next panel to the right). This glass array was interrogated with probes from 64 tissues upon which 114 genes were selected as potential marker genes. Of these, 78 genes were selected for expression validation by RealTime PCR. During the latter, the number of potential marker genes was reduced on the first panel of 81 tissues, and 23 genes were passed on to the second panel which added another 83 tissues. Likewise, 15 genes still displayed a stronger expression in the ovarian tumors compared to the normal tissues and they were tested on an panel of additional 38 tissues. Of these, 5 were selected for further characterization by analysis of protein expression in tissues and patient sera (ELISA). For this, monoclonal have been generated against fusion proteins. In two cases, antibody assays were already in place and patient sera have been screened for the presence of MSLN and SLPI. For WFDC2 two

monoclonal antibodies were generated for use in tan ELISA. Whereas SLPI protein can be found in sera of both patients and controls, MSLN and WFDC2 are found primarily in the patient sera. WFDC2 is not found in sera from healthy controls.

Work accomplished:

Task 1: Generation of representative cDNA arrays from early and late stage ovarian carcinomas

We created four cDNA libraries from pooled tissues (20 pooled fetal ovaries, 4 benign ovarian cystadenomas, 3 normal ovaries, 4 late stage serous ovarian cystadenomas) plus an additional library made from 6 metastatic ovarian carcinomas. For quality control, from each library, 96 clones were randomly chosen. The clones were sequenced and analyzed by similarity analysis against the non-redundant and EST database. The criteria for a satisfactory cDNA library were an average insert size around 1 kb, a low number of mitochondrial and ribosomal sequences, a limited number of clones with no insert, and significant cDNA diversity (Nelson, Ng et al. 1998). Three out of the five libraries fulfilled these criteria. For the fetal and benign libraries, the average insert sizes were considerably below 1 kb. In addition, the number of clones without insert plus the ones with repeats or genomic fragments exceeded one third, a number far too high for consideration to array. Diversity of clones with homology to known genes was similar in all cases. The titer of the three remaining libraries was small which reflects the fact that we chose not to amplify the libraries in order to have a better representation of the lowly expressed clones. We selected 102,680 clones from the three libraries (9,216 from the normal, 17,664 from the late stage and 83,712 from the metastatic library each) and arrayed the colonies onto 32 sets of 5 nylon membranes, each holding 20,536 colonies. The colonies were lysed and the DNA was fixated onto the membranes using a modified Southern blot protocol (after the membranes were placed on filter paper trenched in denaturing and neutralizing solutions, they were dried and subsequently submerged in fat-free milk with 0.5% SDS and 2 x SSC for 2 hours, upon which they were dried and stored at 4°C until usage). As a result of vigorous testing of lysis protocols, this protocol provided the best signal-to-noise ratio for a colony-based membrane hybridization. One set of membranes was hybridized with a probe recognizing the vector portion of each clone. The resulting hybridization pattern revealed that out of the 120,680 colonies that were arrayed, 97,803 actually grew on the membranes. Figure 1 shows a close view on one such membrane where more than 95% of the colonies give a positive signal with the vector probe.

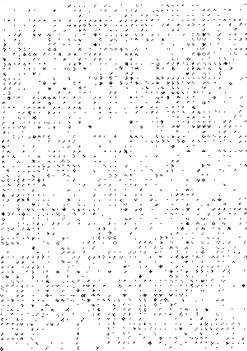


Figure 1 - Sample hybridization

Close view on 1/6 of a membrane containing 3456 colonies that was hybridized with a probe recognizing the vector portion of the cDNA. Where there is no signal, no colony grew. Overall, the number of colonies that did grow reaches 95%

Task 2: Primary characterization of normal and neoplastic tissues using these arrays

The membrane-based cDNA arrays were interrogated with 33P-labeled first-strand cDNA probes which were reverse transcribed using an oligo-dT19V primer from 100 µg of total RNA generated from the tissues listed in Table 2. These tissues had been accrued through the Tissue Collection Core. For most patients and controls, several tissue blocks were generated and some remained in the depository at the Marsha Rivkin Center for later use. Likewise, blood was drawn from each patient and questionnaire data was generated. For details refer to the Core section. The sera were used at the end for the validation of potential serum markers found in the course of this project (Task 5).

Tissue Type	Description
Liver	liver from Clontech
pbl male	white blood cells from 140 ml blood
pbl mix	pooled RNA from 3 controls (male and female)
pbl female	white blood cells from 160 ml blood
pbl culture	lymphocyte culture
normal cyst	paraovarian cyst, 0.95 g
normal ovary	normal ovar./tube tissue, 1.05 g
normal ovary	normal ovar./tube tissue, 0.15 g
normal ovary	normal ovar./tube tissue, right ovary, 1.06 g
normal ovary	normal ovar./tube tissue, left ovary, 270 mg
normal cyst	right ovary, Paraovarian cyst, 720 mg
normal ovary	normal ovar./tube tissue, right ovary, 250 mg
normal ovary	normal ovar./tube tissue, left ovary, 550 mg
normal ovary	normal ovar./tube tissue, right ovary, 520 mg
normal ovary	normal ovar./tube tissue, 0.47 g
normal ovary	normal ovar./tube tissue, left ovary, 0.80 g
normal ovary	fallopian tube from patient with tumor t037
benign ovarian tumor	serous cystadenoma, 0.476 g
benign ovarian tumor	serous cystadenoma, 0.9 g
mucinous stage I	mucinous carcinoma, grade A, stage Ia, 1.6 g
serous stage III	serous carcinoma, grade C, stage IIIC, 0.5g
serous stage III	serous carcinoma, grade B, stage IIIC
serous stage III	serous carcinoma, grade C, stage IIIC, 1.05 g
serous stage III	serous carcinoma, grade C, stage IIIC, 1.54 g
serous stage III	serous carcinoma, 0.92 g
serous stage III	serous carcinoma, grade B, stage IIIC, 0.37 g
undiff. stage III	undifferentiated carcinoma, grade C, stage IIIC, 0.37 g
serous stage III	Serous carcinoma, grade C, stage IIIC, 0.71 g.
serous stage IV	adenocarcinoma, NOS, grade C, stage IVb, 4.3g
serous stage IV	adenocarcinoma, NOS, grade C, stage IVa 1.4 g
serous stage IV	serous carcinoma, grade C, stage IVa 1.25 g
serous stage IV	adenocarcinoma, NOS, grade C, stage IVb, 0.61 g

Table 2 - Tissues used for interrogation of the 100,000 clones membrane array

We used normal, non-ovarian tissues (blue), normal ovarian tissues (green), benign ovarian tumors (orange) and invasive ovarian carcinomas (red)

The probe preparation, hybridization of the membranes and extraction of the hybridization intensities was performed as described earlier (Schummer, Ng et al. 1999). The intensity value for each cDNA hybridized with one of the 30 tissues was stored in a database. This database thus contains the entries from 102,680 cDNAs and 45 hybridization events (30 tissues of which 5 had been hybridized 3 times and 2 twice, plus the hybridization with the vector probe and a "housekeeping" probe recognizing housekeeping genes that are commonly overexpressed in tumors but have no relevance as markers). In addition, the database contains the patient information gathered during tissue accrual by the patient questionnaire, as well as the marker status for ERBB2 (Her2/neu) and TP53 (p53) from the marker tests performed in the core laboratory. All in all the database contains more than 4.8 million entries. Of the 103,680 clones that could have been present on the membranes, 97,802 grew as colonies. This is the number we will further refer to a total number of clones. Once in a digital format, the data was analyzed using the most recently developed algorithms for expression analysis.

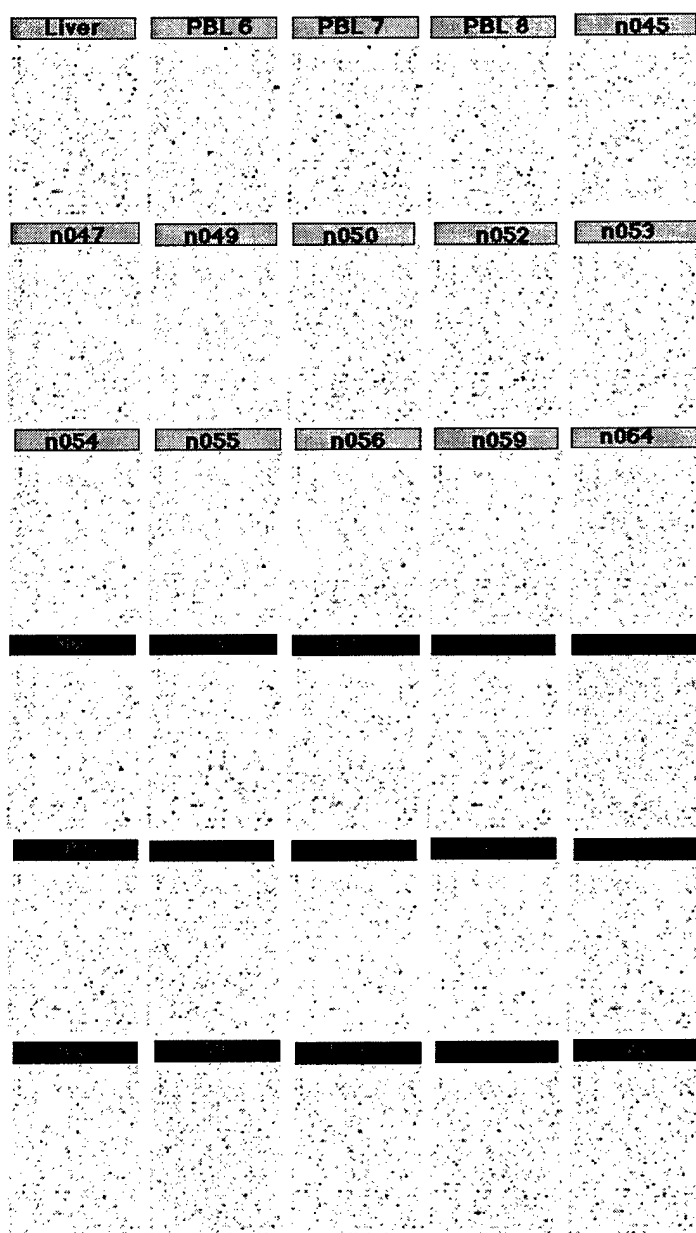


Figure 2 - Hybridization result

Displayed is one field containing 3456 colonies, replicated 30 times and hybridized with probes from 30 different tissues as indicated by the color. Although it may be possible to spot the most obvious differences and similarities in the hybridization pattern by eye, a computer-guided image processing is necessary to detect more subtle changes in expression.

The first task was to identify and exclude from further analysis the clones that code for genes previously known to be overexpressed in cancers due to their higher metabolic rate. We found earlier that these clones are often expressed at high levels (Schummer, Ng et al. 1999). During analysis, their high values would bias the dataset. We designed a probe composed of 41 housekeeping genes (beta actin, comtase, elongation factor 1 alpha, elongation factor 1 gamma, mito-atp6, mito-co1, mito-co2, mito-co3, mito-cyb, mito-nd1, mito-nd2, mito-nd4, mito-nd5, mito-nd6, oviductal-glycoprotein, ribosomal protein L18, ribosomal protein L27,

ribosomal protein L3, ribosomal protein L30, ribosomal protein L5, ribosomal protein L6, ribosomal protein L7, ribosomal protein L7a, ribosomal protein L9, ribosomal protein P0, ribosomal protein S11, ribosomal protein S12, ribosomal protein S13, ribosomal protein S14, ribosomal protein S16, ribosomal protein S17, ribosomal protein S18, ribosomal protein S21, ribosomal protein S24, ribosomal protein S25, ribosomal protein S28, ribosomal protein S3a, ribosomal protein S4, ribosomal protein S6, 18 S rRNA, 28 S rRNA) and hybridized a membrane set with it. This probe will be further referred to as the "junk" probe. The hybridization pattern clearly identified three categories of positive clones with strong, medium-strong and weakly strong signals. We sequenced 30 clones from each category. Only the clones from the two high-expression categories were entirely homologous to the 41 genes in the pool. Therefore we selected only those 10,716 (11%) clones for exclusion from further analysis.

The second step was to reduce the number of clones from the remaining 87,086 clones to 2000-3000, the number that will be arrayed on the second generation cDNA array on glass. Since the goal of our project was to discover genes with potential as markers, preferentially serum-based ones, we focused on the genes with overexpression in the tumors versus the normal tissues. In collaboration with Dr Andy Siegel, who is an adjunct professor of Statistics at the University of Washington we employed statistical measurements to reduce the number of clones in this immense dataset from 87,068 to 2,651. The selected clones exhibit a tendency to a higher expression in the tumor tissues. Table 3 lists the statistical algorithms that were employed for the reduction of the dataset.

		#accumulate to
zScore > 10.09 in >1 of all tumors	1192	1192
† Statistic > 4.00	300	1476
avg.(Tumor) > 2.5* avg.(NormalOvary)	277	1661
avg.(Tumor) > 2.5* avg.(NormalOvary, PBL, liver)	624	2181
avg.(zScore) > 1.4	1439	2949
minus "junk"	298	2651

Table 3 - Clone selection by statistics

Statistical analysis that led to the 2651 selected clones. Each statistical method selected a certain number of clones that added up to 2949. The "junk" probe was a probe consisting of 41 housekeeping genes (ribosomal proteins, mitochondrial genes, elongation factors) that were previously found to have elevated expression in carcinomas presumably due to the elevated metabolism. It reduced the number of clones by 298 to 1651.

We sequenced all 2,651 clones on their 5' ends and the results submitted to homology search in the nr and estdb databases. The result of this homology search is summarized in Table 4. We did not intend to spend too many of our resources on the sequencing. Therefore we opted for a single amplification, single pass sequencing approach. A clone that fails to PCR amplify or that fails to produce a satisfactory sequence would therefore be labeled as "currently unsequenceable", to be attempted to sequence at a later stage. Of our 2,651 clones, 2,061

generated sequences that could be submitted to database homology search. This excludes the clones labeled as "uninformative" in Table 4. Of the remaining clones, 519 were grouped in a class termed "uninteresting", meaning that these genes are known to be expressed at higher levels in cancers because they are either linked to the metabolism (mitochondrial and ribosomal proteins) or expressed in tumor infiltrating lymphocytes (MHC, immunoglobulins). The remaining 1542 informative clones were grouped into those who matched with more than 80% homology to sequences in the nr database ("known"), those who only matched only to sequences in the esdtdb database ("EST") and those who match to neither of the two ("Novel"). In the cases of a hit to only the EST database, we would note how many ESTs our clone was homologous to and the tissues those ESTs were derived from (data not shown). This would indicate whether our clone represented a frequently expressed gene (many hits in the EST database) or, which is more desired, a rare gene, and whether it is found in many tissues or rather in just the ovary.

	#	%	Comment
Total selected clones	2651		
Bad PCR	302		
Sequenced	2349		
Bad sequence	88	4%	
Short sequence	18	0.8%	
Vector	38	2%	
PolyA	107	5%	
Repeat	37	2%	SINE and LINE, genomic, simple repeats
All uninformative	288	12%	
Mitochondrial	203	9%	
Ribosomal protein	19	1%	
Immunoglobulin	310	13%	
MHC	104	4%	
All uninteresting	636	27%	
Novel	45	2%	all unique
EST	298	13%	all unique except for 7 contigs containing 17 clones
Full length EST	68	3%	
All ESTs	366	16%	
GAPDH	142	6%	
Ferritin H	84	4%	
IGF-2	63	3%	
collagen 1A1	32	1.4%	
SLPI	30	1.3%	
S100A6	18	0.8%	
HE4	17	0.7%	
S100A11	10	0.4%	

Others	618	26%	
All known genes	1014	43%	excluding Ig, MHC, Mito, repeats, vector
All unique genes	883	38%	comprising 467 Known genes, 366 ESTs and 45 Novels

Table 4 - Identity of potential marker genes

Explanation of the terms used: "bad PCR" means that the PCR amplification of the clone resulted no band, multiple bands or a smear, we did not attempt to repeat the reaction; "bad sequence" refers to a sequence with an unreadable chromatogram, the sequencing reaction was not repeated; "EST" refers to clones that have no hit in the nr database at GenBank but one or several hits in the estdb; "full length EST" refers to the clones resulting from the full-length sequencing projects (KIAA, DKFZ, FLJ etc.); "known genes" refers to clones that have a hit of 80% or more in the nr database at GenBank.

Our libraries were all oligo-dT primed and hence the clones should all represent the 3' ends of transcripts. All of the novel genes represent unique sequences which means that we have identified 45 novel sequences with potential elevated expression in the ovarian carcinomas. Of the ESTs, 17 clones could be matched to 7 contigs. Therefore the 366 ESTs correspond to a maximum of 356 genes. Of the clones matching to known genes, the eight genes with the most clones representing them are listed in Table 4. They are discussed briefly.

GAPD, or glyceraldehyde-3-phosphate dehydrogenase, was the most abundant one with 6% of all sequenced clones. The increased expression of GAPD in cancers was reported earlier (Tokunaga, Nakamura et al. 1987; Schek, Hall et al. 1988; Persons, Schek et al. 1989; Desprez, Poujol et al. 1992; Finnegan, Goepel et al. 1993; Chang, Juan et al. 1998; Kim, Kim et al. 1998). In the past, GAPDH has been used for normalization of Northern blots which speaks for its ubiquitous expression.

FTH1, or ferritin H, transcript was found to be elevated in ovarian tumors (Tripathi and Chatterjee 1996). Ferritin serum levels have been reported to be elevated in ovarian cancer patients (Yuan, Ng et al. 1988; Lahousen, Stettner et al. 1989; Pinto, Marinaccio et al. 1997). However, due to its low specificity, Ferritin is not suited as a diagnostic marker (Pinto, Marinaccio et al. 1997).

IGF2, or insulin-like growth factor 2, was reported to play a role in ovarian cancer. In some carcinomas, IGF-2 loses its chromosomal imprinting resulting in an overexpression (Yun, Fukumoto et al. 1996; Chen, Ip et al. 2000). Antisense oligonucleotides against IGF-2 inhibited cell proliferation and induced apoptosis in human ovarian cancer AO cells (Yin, Pu et al. 1998).

Collagens of classes 1 and 3 were reported to be actively produced both locally in the ovary as well as more remotely in the peritoneal cavity (Kauppila, Saarela et al. 1996). Collagens and procollagen serum levels may be indicative of ovarian cancer disease outcome (Santala, Simojoki et al. 1999).

WFDC2, or HE4, is a secreted protease inhibitor previously found by us to display elevated transcript levels in ovarian carcinomas (Schummer, Ng et al. 1999). It is regarded as our gold standard, meaning that the selection of potential marker genes should contain this gene, otherwise our selection criteria may need revision.

SLPI is also a secreted protease inhibitor, albeit with a different sequence. It is expressed by several glandular epithelial cells and it is thought to have anti-bacterial anti-HIV properties (Wingens, van Bergen et al. 1998). There is a SLPI ELISA test commercially available which will be described further below. SLPI has not been implicated in any cancer.

S100A6, or prolactin receptor-associated protein PRA, or calcyclin, binds GAPD (Filipek, Wojda et al. 1995). Its protein is overexpressed in a variety of tumors including colorectal adenocarcinomas (Komatsu, Andoh et al. 2000).

S100A11, or calgizzarin, or S100 calcium-binding protein A11, is expressed in colorectal carcinomas (Tanaka, Adzuma et al. 1995) and may be involved in the regulation of cell transformation and/or differentiation (Moog-Lutz, Bouillet et al. 1995). Genes of the S100 family are implicated in a variety of cancers, among them melanoma (Van Ginkel, Gee et al. 1998) and breast (Pedrocchi, Schafer et al. 1994).

The aforementioned 1542 informative clones correspond to 883 unique genes. They have a high potential to be marker genes, but as pointed out earlier, there is a significant error associated with them and a large portion of them may have been selected as false positives. Before we can validate the expression of these genes by a method different from array hybridization (namely by quantitative RealTime PCR), we will narrow down their number. This will be achieved in Task 3.

Task 2 (addendum): Cluster analysis of 2651 clones

The accrual of ovarian cancer tissues, especially the early stage serous carcinomas) was less effective than originally anticipated. With the end of Task 2, the few early stage cancer tissues that the tissue collection core was able to accumulate were of non-serous histology. Therefore we postponed the beginning of Task 3 (further characterization of gene expression by glass array) and instead attempted to analyze the data generated by the membrane array. This data was never meant to be analyzed in depth due to the high variability associated with it. We reasoned that in spite of the shortcomings, the genes with the most striking tumor-typical expression would stick out. The facts that we based our assumptions on are as follows.

1. The array contains genes from ovarian tumor libraries, hence a gene with high expression in the tumor should be present with multiple copies. Chances are that only a few copies of a gene show a suboptimal hybridization result and that the other copies can be used for proper analysis.
2. For some genes that were represented by multiple clones (such as for WFDC2, SLPI, GAPD etc.), we calculated the standard of means as an assessment of expression variation between these clones. When calculated for each tissue separately, the standard of means

averaged at 75%, ranging from 20% to 190%. When we averaged for each clone its expression values across the tumors and, separately, across the normal tissues, followed by a calculation of the two standard of means, they averaged at 23%, ranging from 10% to 37%. This is a significant reduction in variability, which leads us to the next conclusion.

3. Knowing that each individual clone is associated with high error, rather than looking at the expression of one gene across all tissues we would use an analysis tool which takes into account the expression of all genes in all tissues simultaneously. This software, Bioclust™ (Ben-Dor et al., 2000b), was designed for large expression data sets like the present one. In its initial stage it was capable of analyzing datasets of maximal 5000 clones within reasonable time limits. The updated version that is available today has no such limitations.

The dataset generated in Task 2 consisted of 2651 clones assayed on 30 tissues (or hybridizations). In order to provide an assessment of the variability between hybridizations, we selected 7 tissues and repeated their hybridizations twice more (on 5 tissues, resulting in triplicated hybridization) or once more (on 2 tissues, resulting in duplicated hybridization). This resulted in an extended dataset with the same number of clones but 16 more hybridizations.

In order to determine the degree of consistence between hybridizations of the same probe, the standard of means of each clone across the two or three repeated hybridizations was calculated and averaged. The average standard of means for the replicate hybridizations is 61%, ranging from 6% to 161%. For comparison, the same value for 8 hybridizations with 8 probes from 8 different tissues averaged at 71%, ranging from 10% to 210%. Viewed in the context of the vast majority of the clones (~95%) expressing uniformly across all tissues (Schummer, Ng et al. 1999), this shift in the standard of means is significant. In addition, as will be shown below, when treated as hybridization probes coming from separate tissues, the replicates display a higher tendency to cluster together than the unrelated tissues.

We performed two separate clusterings, the ones of the tissues and the ones of the clones (Figure 3). In both cases, the algorithm was performed several times with varying starting parameters and varying fractions of the dataset. Please refer to our recent publication for details (Ben-Dor, Bruhn et al. 2000).

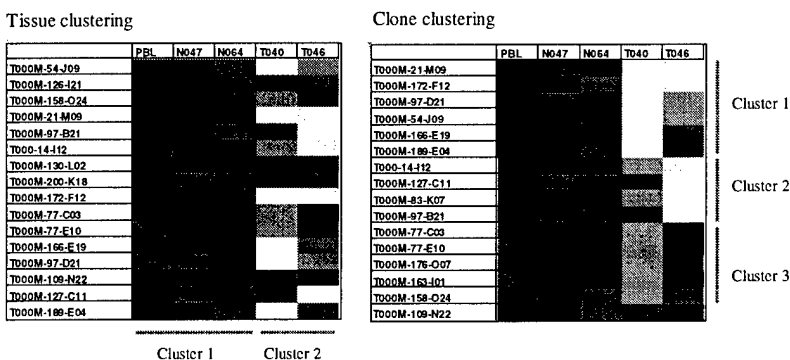


Figure 3 - Schematic explanation of the clustering

For better visual impression, the dataset is represented as a table and the values have been replaced by greyscale where white stands for high expression. Shown are 16 clones out of the 2651 (in the rows) and 5 hybridizations out of the 46 (in the columns): PBL (peripheral blood lymphocytes), two normal ovaries (N...) and two ovarian tumors (T...). In the left panel the tissues were clustered into two groups, one consisting of the normal ovaries and the PBL, the other consisting of the tumors. In order to select potential marker genes, the same clustering algorithm was repeated with a decreasing number of clones that would sort the tissues as nicely as displayed. The minimal number of clones that achieve this grouping are regarded as potential markers. In the right panel the clones were clustered into three groups. It is conceivable that members of a group are either clones representing the same gene or gene family or genes that share similar function or similar pathways. A clone that consistently clusters with a known tumor gene would be regarded as a potential marker gene. The small example shown here was applied to the full dataset as shown in Figure 4.

The clustering of the clones grouped certain clones together that turned out to be either mostly copies of the same gene or genes with similar behavior (Figure 4). One such group contains the "gold standard" WFDC2, or HE4, that was found earlier to be a potential marker gene for ovarian cancer (Schummer, Ng et al. 1999), together with other genes, among them SLPI, a secreted protease inhibitor just like WFDC2. Other clones from this group show no match to any known gene and may be potential novel marker genes.

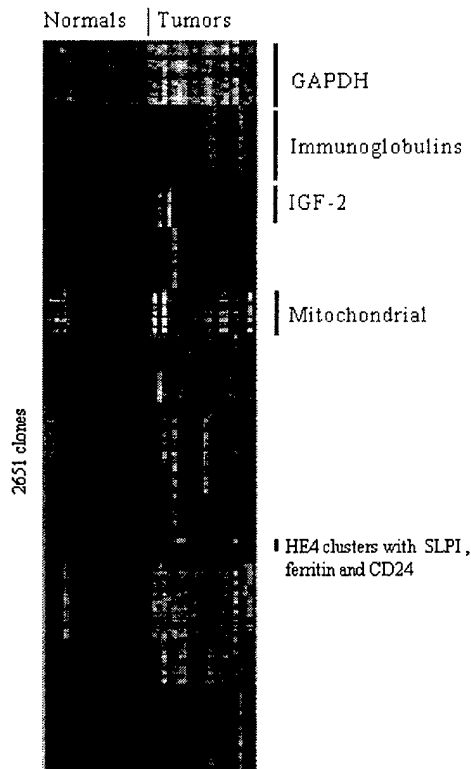


Figure 4 - Clone clustering on full dataset

Clone clustering performed on the full dataset of 2651 clones. The expression values are displayed as greyscale with white standing for high expression and black for a low one. The normal tissues (liver, PBL, normal ovaries) are shown on the left, the ovarian tumors on the right. Overall the expression of the normal tissues is lower than that of the tumors which reflects the selection criteria of these 2651 clones (low expression in normal tissues, high in tumors). In the present example the clones were clustered into 75 groups of varying size. The biggest groups consist to more than 80% of clones matching to GAPDH, immunoglobulins, IGF2 and mitochondrial genes. Some of the smaller groups contain known tumor genes (such as CD24, FTH1 and WFDC2) together with genes that were previously not known to be associated with tumors (such as SLPI and clones that do not match known sequences in the public databases). These clones were regarded as potential marker genes.

The clustering of the tissues was initially performed on the full dataset. The specific clustering experiment that was performed was of the leave-one-out nature. Briefly, all tissues were labeled as either normal, tumor or none of the two. The clustering was performed using all tissues but one. Bioclust would determine which clones were particularly useful to achieve the best separation between the "tumor" and "normal" groups. Based on these clones, the left-out tissue was introduced into the analysis and it was recorded whether it would classify right (e.g. a tumor classifying as a tumor) or wrong (e.g. a normal classifying as a tumor). This experiment was reiterated until all tissues had been left out once. This experiment was performed several times with varying starting parameters. An example of a typical set of clones that resulted from this analysis is shown in Table 5. An example of an optimal tissue clustering result is given in Figure 5.

#	Gene name	Comment
62	GAPDH	ovarian cancer
6	ferritin H	
4	collagen 1A1	
3	Immunoglobulin	
2	EST	TIL
2	HE4	ovarian cancer
2	Keratin 18	breast cancer
2	MHC	TIL
2	Mitochondrial	metabolism
2	Novel	
2	SLPI	
1	TMPO	cell proliferation
1	SSR4	
1	PKM 2	hepatoma
1	Lactate dehydrogenase	
1	COX7b	
94	TOTAL	

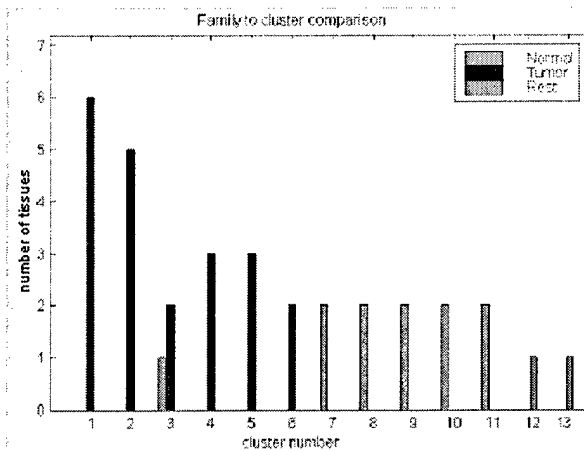


Figure 5 - Example of a tissue clustering result on the entire dataset

Displayed is a typical result for the leave-one-out tissue clustering analysis. The software generated 6 groups which - with the exception of one normal tissue - consist of tumors and five groups that contain only normal tissues. The duplicate and triplicate hybridizations of one tissue were treated as if they had been derived from separate tissues. As a result they either cluster in separate groups, which would be an indicator of low similarity, or they cluster in the same groups, indicating that they are indeed very similar to each other. Of the 7 tissues with repeated hybridizations, 5 have their replicates cluster in the same groups, one has two replicates in a "tumor" group and another replicates in a neighboring "tumor" group, and one has two replicates in a "normal" group and a single replicate in a "tumor" group. The groups 1-13 are formed from the following tissues: 1: hwbc3, t037, t051, t051a, t040, t065; 2: t025, t060, t066, t044a, t044b; 3: n050a, t048, t044; 4: t046, t046a, t046b; 5: t063, t048a, t048b; 6: n039a, t043; 7: hpbl7, hpbl8; 8: n047a, n047b; 9: n050, n050b; 10: hliv2, hpbl6; 11: n056, n064; 12: t062; 13: t058. An "a" or a "b" behind the tissue name refers to the duplicate and triplicate hybridization.

Each of the clustering experiments resulted in a list of genes (See legend to Table 5) of which many were found to be the same in different experiments. These clones included such metabolism-related genes as the mitochondrial genes, ribosomal proteins, elongation factors and GAPDH. The non-metabolism genes which were picked up by all clustering experiments are listed in Table 6. WFDC2, SLPI and S100A11 have been discussed above with respect to their tumor-relatedness. ACTB, or beta actin, transcript was reported to be expressed at higher levels in colorectal neoplasia (Naylor, Stamp et al. 1992). CD24 is a known marker for breast cancer (Fogel, Friederichs et al. 1999). ELF3, or ESE1, is an epithelial-specific transcription factor (Oettgen, Alani et al. 1997) related to the ets family and is expressed in lung carcinomas (Tymms, Ng et al. 1997). FOLR1 was previously reported to be overexpressed in ovarian carcinomas (Toffoli, Cernigoi et al. 1997). GPR39 is a G-protein coupled receptor (McKee, Tan et al. 1997). KRT8, or keratin 8, is an epithelial gene that was reported to be expressed in a variety of tumors. It may be of diagnostic value in cervical cancer (Martens, Baars et al. 1999). PAX2 is expressed in Wilm's tumors (Davies, Perera et al. 1999). It encodes a DNA binding, transcription factor whose expression is essential for the development of the renal epithelium (Dressler and Woolf 1999).

This selection of genes showed that the cluster analysis was capable of detecting among our 2651 clones a large number of cancer-related genes. It was therefore our primary interest to a) confirm their expression by a method other than array hybridization and to focus on the genes and clones with no previous cancer role ascribed to them, such as SLPI and GPR39 and the sequences with no match to the known gene databases. The validation of gene expression is described in Task 4.

Gene name	GenBank Accession Number
5 ESTs	AA522512, AI271417, AA131674, AW300236, AL080004
2 novel sequences	
beta-actin	NM_001101
CD24	L33930
ESE-1	U73844
Folate BP	X69516
GPR39	AF034633
HE4	X63187
Keratin 8	G4504918
PAX 2	AH006910
S100A11	D38583
SLPI	NM_003064

Table 6 - Selection of genes that were found by the cluster analysis of the membrane data based on the expression of 2651 clones in 30 tissues

After performing several rounds of cluster analysis both of the clones and the tissues, we found more than 100 clones that were positive in all experiments. Of these, most coded for metabolism-related genes, including GAPD, with low marker potential. The other genes are listed here.

Task 3: Further characterization of gene expression in normal and neoplastic ovarian tissue

The 883 genes identified in Task 2 should ideally display an expression pattern that is higher in the tumors than in the normal ovarian tissues. There are several reasons why this observed behavior may not coincide with the actual one. Firstly, there were only 32 tissues used and we didn't know how a gene would fare in other tissues. Secondly, there was heterogeneity in cell composition between tissues of the same kind and within the same tissues. Thirdly, the array consisted of single-spotted colonies, and commonly triplicate spotting and above is regarded as statistically relevant (Geiss, Bumgarner et al. 2000; Ichikawa, Norris et al. 2000). Fourthly, the method of hybridization, and image processing adds a certain variation to the values. As a consequence, we did not regard the 883 genes as the final set of cancer genes and proceeded to reanalyze their expression on glass arrays hybridized with the same tissues as before and with additional tissues (Table 7). Glass arrays, if processed properly, have lower signal-to-noise ratio than membrane arrays and due to double spotting combined with double hybridization, each value is more dependable. But even here some of the aforementioned factors apply.

For each of the 883 genes we selected the longest clones and the ones with the best sequence. Our cDNA glass array could hold as much as 1536 genes or clones. Some of these positions are reserved by controls such as RNA, polyA, non-human clones (Arabidopsis), vector sequence and repeats. These controls amounted to 23 positions, leaving us with 1513 positions to fill. From earlier cDNA expression arrays, we had accumulated clones with potential as markers for ovarian cancer, one of them being the dataset published in Gene (Schummer et al., 1999). These genes together with our 883 genes were PCR amplified using vector-specific primers (mapping 150 bp upstream and downstream from the multiple cloning site) in 100 µl reactions. The PCR products were concentrated to 10 µl, and 10 µl of 100% DMSO was added to prevent evaporation during the arraying process. We used a Generation II arrayer from Molecular Dynamics to array the cDNAs onto Type 7 "mirrored" slides from Amersham. These slides contain an aluminum coating rendering them reflective, thus maximizing the photon yield. Each cDNA was spotted as duplicate. The arrays were hybridized with first-strand cDNA probes generated from the tissues listed in Table 7. The hybridization and data extraction was performed according to the conditions described earlier (Geiss, Bumgarner et al. 2000) with the exception of the reference probe which was generated from a pool of RNAs from all 64 tissues. Glass arrays allow for cohybridization of two probes, one labeled with Cy3 and one labeled with the Cy5 dye (further referred to as green and red dye). The red color is used on the tissue to be interrogated, the green on the reference. The data was processed by software developed by Roger Bumgarner at the University of Washington which merges the duplicate expression values and their local background values into one averaged value and an error assessment value. These values were written to a database. As a result, for each gene on the array and for each tissue, the database hosted a record of the expression relative to the reference.

Tissue Type	Tissue Type
* Liver	
* white blood cell culture	
* normal ovary	* stage I ovarian carcinoma
* normal ovary	* stage III ovarian carcinoma
* normal ovary	* stage III ovarian carcinoma
* normal ovary	* stage III ovarian carcinoma
* normal ovary	* stage III ovarian carcinoma
* normal ovary	* stage III ovarian carcinoma
* normal ovary	* stage III ovarian carcinoma
* normal ovary	* stage III ovarian carcinoma
* normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma

normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
normal ovary	stage III ovarian carcinoma
fallopian tube from ovarian cancer patient	stage III ovarian carcinoma
* benign ovarian cystadenoma	stage III ovarian carcinoma
* benign ovarian cystadenoma	stage III ovarian carcinoma
benign ovarian cystadenoma	* stage IV ovarian carcinoma
benign ovarian cystadenoma	* stage IV ovarian carcinoma
benign ovarian cystadenoma	* stage IV ovarian carcinoma
benign ovarian cystadenoma	* stage IV ovarian carcinoma
benign ovarian cystadenoma	stage IV ovarian carcinoma

Table 7 - Tissues used for interrogation of the 1536 clones glass array

The 64 tissues are color coded. Blue stands for normal, non-ovarian tissues, green for normal ovarian tissues, orange for benign ovarian tumors and red for invasive ovarian carcinomas. The asterisk in the first column marks the 25 tissues that were previously used for the interrogation of the membrane array. For the remaining 5 tissues from the membrane array, the tissue RNA was used up for the interrogation of the membrane array and the experiment could not be repeated on glass.

Data analysis

During the last 4 years, algorithms had been developed for the analysis of complex biological datasets. Some were originally designed to sort and understand other scientific datasets, such as those generated in epidemiology, others were tailored to the array data. We used the clustering approach to extract from our data the genes with potential for markers. These genes needed to show higher expression in the ovarian tumors (invasive, benign or both) than in the normal ovaries and in the liver. The algorithm we employed was written for array data and optimized during our month-long analysis.

We performed clustering analysis using Bioclust" (Ben-Dor, Bruhn et al.2000), clustering both tissues and genes. The clustering experiments were performed as described above under "Task 2 addendum". We found 126 genes with elevated expression patterns in the tumors, among them 8 novel genes and 30 ESTs. These genes are listed in Table 8. It is remarkable that this list encompasses the genes found earlier when analyzing the membrane data (see Table 6), which is another proof of the quality of our analysis tools given the high degree of error associated with the membrane data. It is also remarkable that all 7 marker genes found earlier to be overexpressed in ovarian tumors by screening an array of 21,00 cDNAs with probes from five ovarian tumors and OSE are contained in this dataset (Schummer, Ng et al. 1999). These genes are YWAHQ, MFGE8, E16, WFDC2, MUC1, the putative progesterone binding protein and SDC4.

Gene name	GenBank accession	Gene name	GenBank accession
14.3.3	x56468	IFI27, interferon-induced protein 27	NM_005532
actin, beta	NM_001101	IGF BP2	X16302
adenocarcinoma-associated antigen (KSA)	X14758	IGF2	X07868
amyloid protein homologue	L09209	Kadereit	L22343
argininosuccinate synthetase (ASS)	NM_000050	Keratin 18	NM_000224
BA46	U58516	Keratin 7	NM_005556
bamacan	AF067163	Keratin 8	G4504918
bikunin	U78095	KIAA0762	AB018305
c-jun	NM_002228	LDHA, lactate dehydrogenase A	NM_005566
c-myc	X00364	LGALS1, lectin, galactoside-binding, soluble	NM_002305
Calvasculin	NM_002961	lipocalin 2 (oncogene 24p3) (LCN2)	NM_005564
CCR2	D29984	Lipocalin2	NM_005564
CD24 signal transducer	L33930	MAGE-4	D32075
CD9 antigen (p24)	NM_001769	MAGOH, mago nashi homolog	AF035940
CDC28 protein kinase 2 (CKS2)	NM_001827	MAT1	L37385
CGGBP, trinucleotide repeat DNA BP p20-CGGBP	AF094481	MCAF	M24545
chaperonin	X74801	MDC15	U46005
CHI3L1, chitinase 3-like 1 (cartilage glycoprotein-39)	NM_001276	Mesothelin	AF180951
collagen 11A1	NM_001854	MET	NM_000245
CRIP1, cysteine-rich protein 1 (intestinal)	NM_001311	MIS	K03474
cyclin-selective ubiquitin carrier protein	U73379	Mucin1	X52228
cytosolic malate dehydrogenase	D55654	NME4	NM_005009
DAP-1 (ST kinase)	X76105	OGP, oviductal glycoprotein exon 11	U58010
density-regulated protein (DRP)	NM_003677	Osteopontin	D14813
E16	M80244	oviductal glycoprotein	U09550
EDN1	S56805	p27 alpha-inducible protein 27 (IFI27)	X67325
Efs1 or Efs2	AB001466	p73	NM_005427
Enolase	NM_001428	p76, endosomal, multispinning membrane prt.	U81006
ESE-1	U73844	Pax2	AH006910
FACL3, fatty-acid-Coenzyme A ligase, long-chain 3	NM_004457	PLTP	NM_006227
Ferritin H	L20941	progesterone binding protein	Y12711
Folate BP	X69516	pyruvate kinase, muscle (PKM2)	NM_002654
GA733-1	NM_002353	RIG-E	Z68179
GAB2	AB018413	Ryudocan	D13292
GAPDH	M33197	S100A11	D38583
glia maturation factor-gamma (GMF-GAMMA)	NM_004877	SAS	U01160
GPR39	AF034633	SCNN1A, sodium channel, nonvoltage-gated 1 alpha	NM_001038
gpx1, glutathione peroxidase	X13709	SLPIa	NM_003064
haptoglobin	NM_005143	ST5	NM_005418
HE4	X63187	STK11	AF035625
HSPD1, heat shock 60kD protein 1	NM_002156	TPI1 triosephosphate isomerase	M10036
Her2/neu	M11730	tra1, homolog of murine tumor rejection antigen gp96	X15187
HGF	X16323	TAGLN2, transgelin 2	NM_003564
HSD3B2	M77144	UNC119	AF125997

Table 8 List of 88 genes (excluding the 8 novel genes and the 30 ESTs) found as a result of the cluster analysis of the glass array data set. All genes are listed with their GenBank accession number for easy identification.

Array hybridization is ideal for the determination of the expression of thousands of genes in dozens of tissues. The array method, however, even the glass-based method using duplicate spotting, has several instances where error is introduced. Firstly, the RNA quantification by spectrophotometry is inaccurate, with an standard of means of 30-50% (unpublished results). Secondly, first strand cDNA generation and probe hybridization do not always use highly reproducible results, even with the highest care taken. Thirdly, the spotting of the DNA onto the glass or the membrane can result in differences of amount of DNA that actually remains on the surface. Fourthly, the method used for spot detection and intensity integration adds a minor but detectable variability to the numbers. Taken altogether, for a single gene, we estimated the average error to be as high as 50% of its measured intensity. For this reason, we validated the expression of the 126 genes by a method that is a) more accurate than array hybridization, b) capable of processing this large number of clones within reasonable time, and c) more sensitive than array hybridization. The method of choice was quantitative Real-Time PCR and will be described in Task 4.

Task 4: Characterization of highly expressed genes associated with cancer

Task 3 left us with 124 genes to be further characterized. In addition, after discussion with colleagues, we decided to validate 6 more genes with known or suspected involvement in ovarian cancer (BRCA1, AF005068; BRCA2, U43746; ESR1, X03635; ESR2, AB006589; TP53, NM_000546; StAR, U17280) and 11 genes our collaborators within the ovarian SPORES were working on (AKT1, M63167; AKT2, M77198; Cyclin C, M74091; IL-8, M17017; Ku70, J04607; Ku80, M30938; LotI, U72621; MR, NM_013404; NY-ESO-1, U87459; PIK3CA, Z29090; PTEN, U93051). This increased the number of genes to characterize to 141.

As pointed out above, the first step in the characterization of our potential marker genes was the expression validation by means of RealTime quantitative PCR. This method requires the design of two primers per gene, spaced by ~500 bp. The primers need to have similar melting temperatures (T_m) and should all be of 20-23 nucleotides in length. The primers needed to be tested on cDNA that was reverse transcribed from pooled RNA from a number of ovarian cancer tissues (to minimize chances of a negative result due to absence of the transcript in a given tissue). The primer pair will then be used in a conventional PCR supplemented with a fluorescent dye (SYBR green). This dye emits light upon UV excitation in the presence of double-stranded and single-stranded DNA, the latter with less efficiency. The PCR is performed in a 96-well plate (60 s at 94°, 40 cycles of 25 s at 94°, 25 s at 60°, 45 s at 72° using 1 U/μl of Biolase enzyme made by Bioline and 0.12 mM dNTPs, 0.12 mM of each primer, 1.5 mM MgCl₂ and the supplied buffer) in an ABI7700 RealTime PCR machine in which the SYBR green emission is recorded several times during each cycle, thus monitoring in real time the built-up of newly synthesized DNA molecules. The PCR machine comes with software that uses a standard on each 96-well plate to determine the DNA concentration. This

standard consists of a twofold serial dilution of cDNA made from a white blood cell RNA preparation that is amplified using the primers for TMP21 (GenBank accession number U61734), a gene which we find to be expressed in all tissues tested so far.

Since SYBR green cannot distinguish between the actual PCR product and DNA molecules that are made at random (artifacts), the PCR was run on a 1% agarose gel for determination of the quality of the PCR band. In the case of the absence of a band on the gel but the presence of a SYBR signal, we would set the resulting DNA concentration to 0. A typical RealTime PCR result is shown in Figure 6.

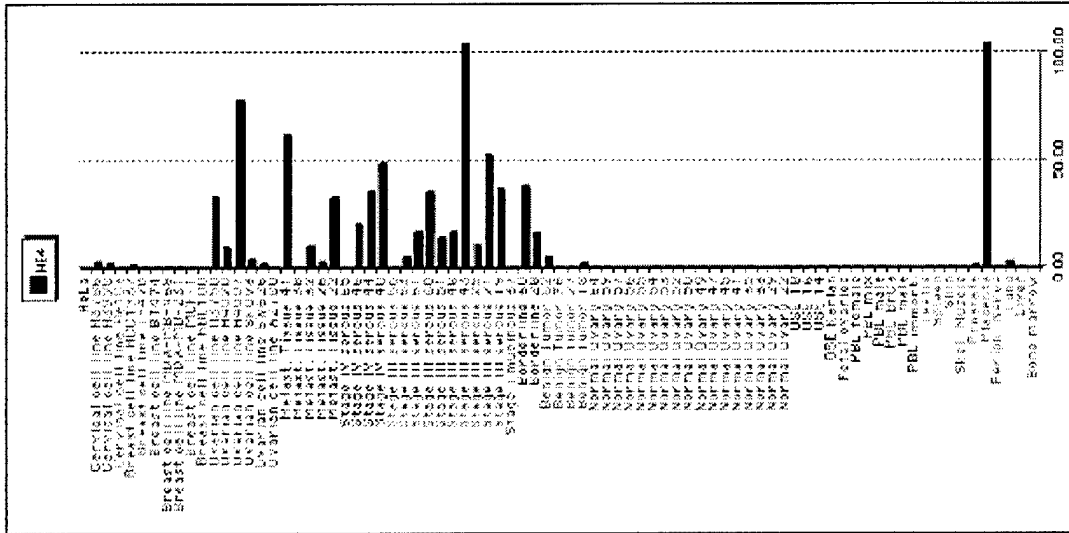


Figure 6 - RealTime quantitative PCR result of WFDC2 on 82 tissues

The tissue names are listed on the bottom. Brown stands for normal non-ovarian tissues, blue for peripheral blood lymphocytes, green for normal ovaries, orange for benign ovarian tumors, red for ovarian carcinomas of increasing stage, and the leftmost 17 entries are ovarian, breast and cervical cell lines. The y-axis shows expression of WFDC2 relative to the TMP21 standard. These are arbitrary values that can nevertheless be used for comparison of the degree of expression of different genes. ACTB, or beta actin, a medium high expressed gene, would show numbers in the 400 range, a lowly expressed gene would show numbers in the 0.1 range. WFDC2 transcript expression is, with the exception of placenta and lung, clearly restricted to the ovarian tumors. This pattern shows WFDC2 as a marker gene with high specificity and sensitivity.

Key to the success of a PCR is the proper design of the primer pair which requires, amongst others, an error-free DNA sequence. While this requirement was met by none of the sequences we have produced (single pass sequencing rarely results in an error-free sequence), we could, in the case of the clones that match to known genes, use the published sequences as template for primer design. However there were caveats that prohibit the generation of functioning primer pairs. In these cases, we generated up to two primer pairs (resulting in four possible PCR products) for one gene before abandoning the primer generation altogether.

In the case of the 30 clones that match only to ESTs (that are also derived by single pass sequencing) the databases gave us several homologous sequences which we compared against

each other, and we would design the primers in regions that were 100% identical. However, one shortcoming of the ESTs is their length. As pointed out above, our PCR products are typically ~500 bp long, but our system was able to handle lengths of 350 and below. The EST sequences are often derived from oligo-dT primed cDNA libraries in which case they cover the 3' untranslated region of their gene. This region often contains repeats such as LINEs and SINEs (50-150 bp length) which are unsuitable for primer placement. In an average 400 bp EST, this may leave less than 300 nucleotides for the placement of the primers, and combined with the possibility of inaccuracy of some base readings, it will be difficult to generate a PCR product. We were therefore unable to generate PCR primers for 8 of the 30 ESTs.

In the case of the 8 clones that did not match to any published sequence all of their sequences were between 250 and 450 bp long with the number of undecided base pairs growing at the 3' end. We were therefore unable to generate PCR primers in 7 cases.

In the case of the 105 known genes (88 from the glass array plus 17 genes suggested by our collaborators) the chances of generating functional primer pairs are very high.

In summary, we could generate functioning PCR primers on 78 genes (see Figure 1).

The RealTime quantitative PCR was performed on three separate 96-well plates containing the cDNA templates. Each cDNA was reverse transcribed from 10 µl of total RNA using the Superscript system (Life Technologies) and after completion water was added to each cDNA preparation to 500 µl. The cDNAs from the template tissues (normal, non-ovarian tissues; normal ovaries; ovarian surface epithelium primary cultures; benign ovarian tumors; borderline ovarian tumors; stage I, II and IV and metastatic ovarian carcinomas; ovarian, breast, cervical cell lines) were transferred into the wells of three 96-well plates.

Plate 1 contained samples from all tissue classes and served as a prescreen. Genes that showed high expression in the normal ovaries or in the normal, non-ovarian tissues were eliminated from further validation (Table 1). This reduced the number of genes by 55. The 23 genes that showed a tendency to express higher in the ovarian tumors than in the normal tissues were passed on to Plate 2 which contained more tissues from each class, especially more ovarian tissues (normal and cancers). Again, genes which failed to show overexpression in the ovarian carcinomas were eliminated. This reduced the number of genes by 8, leaving 15 to be assayed on Plate 3 which contained a large number of normal-non ovarian tissues and which was therefore used as a screen for genes that do not significantly express in these tissues.

Of these 15 genes, 5 displayed a clear ovarian cancer-related expression. These genes were ELF3, GPR39, LCN2, WFDC2, MSLN and SLPI. Figure 7 shows their expression and that of other genes across all 202 tissues assayed.

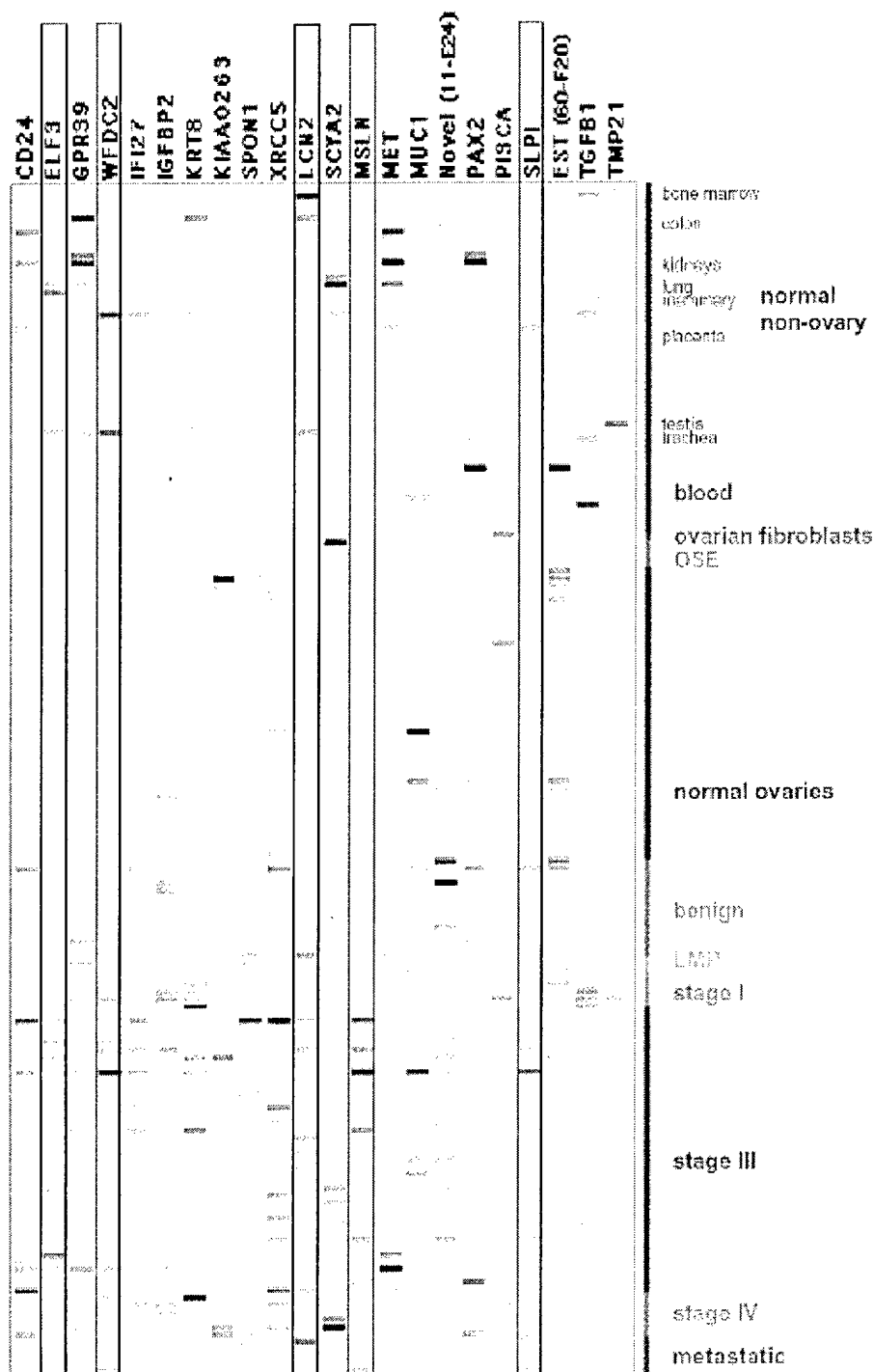


Figure 7 - RealTime data focusing on the expression of the marker genes in all tissues
The expression of 21 genes in 202 tissues was determined by RealTime quantitative PCR. Listed on the right are the tissues using the same colors employed throughout the report. The names of the genes are listed at the top. The expression values are expressed as greyscale bands with black standing for high expression and white for low. The four best performing genes are highlighted. The values are not normalized since normalization requires a gene or

a group of genes with prior knowledge of their unchanged expression in the tissues tested. Since this is impossible, we have included in this panel the gene TMP21 which is expressed in all tissues shown, albeit with some variation. We would like to point out that had we normalized by the values of this gene, the overall expression pattern would still look the same with some bands being darker or lighter than otherwise.

Of the 21 genes whose expression was assayed using real-time PCR more than half are capable to discriminate between ovarian cancers and normal ovaries, some better, some worse, and they are thus potential markers for ovarian cancer. The evaluation of these markers requires the availability of tissue which forfeits a role in early detection screening. Early detection screening requires a non-invasive test which by all standards means that the protein be present in the blood (see below). Nevertheless, the marker genes we have found so far may be useful for staging and prognosis. This needs to be evaluated on a larger set of cancer tissues, not only of the serous histology but also of mucinous and endometrioid. We will then be able to answer questions about the discrimination between benign, borderline and invasive tumors.

Task 5: Final Analyses

The genes found in Task 4 show a great marker potential but so far they require to be tested on tissues which require an invasive procedure to obtain. Biopsies may be acceptable procedure in high risk populations but they are not acceptable for the screening of a general population. Due to the relatively low incidence of ovarian cancer (25,000 new cases every year in the US (American Cancer Society 1998)) early detection can only be achieved using an inexpensive test with high specificity (Urban 1999) and sensitivity that uses body fluids such as blood, saliva or urine. The most commonly employed body fluid-based tests are ELISA which detects proteins via an antibody, Western blot, using an antibody too, and quantitative PCR which detects transcripts in circulating cells. The latter can be done using the information and the resources gathered so far, the protein detection assays require that we express the protein and raise monoclonal antibodies to it.

It has to be pointed out that the transcript level of a gene does not always correlate with the amount of protein that is made (Anderson and Seilhamer 1997). And even if the elevated transcript level were translated into elevated protein level, we estimate the odds to find the protein in the sera of patients to be less than 50%.

We have therefore decided to follow a two-pronged approach, attempting to detect transcripts in circulating cells in blood and peritoneal washes while expressing fusion proteins of the genes and raising monoclonal antibodies against them with the goal of developing a sandwich-ELISA.

We have chosen 5 genes for initial antibody generation: ELF3, GPR39, WFDC2, MSLN and SLPI. ELISA tests are available for MSLN and SLPI proteins. For WDC2 a sandwich ELISA has been designed by Ingegerd Hellström and collaborators at the Pacific Northwest Research Institute in Seattle. For GPR39, we are currently expressing fusion proteins (in collaboration

with Drs Ingegerd and Karl Erik Hellström and Dr Jeff Ledbetter at the Pacific Northwest Research Institute in Seattle) for the immunization of mice which will ultimately lead to the generation of monoclonal antibodies.

SLPI protein in serum

SLPI codes for a secreted protease inhibitor that is expressed in mast cells where it may inhibit a mast cell chymase (Westin, Polling et al. 1999). SLPI inhibits leukocyte-derived proteinases, has anti-HIV-1, antibacterial, and antifungal properties, and interferes with the induction of synthesis of proinflammatory mediators in monocytes and macrophages (Wingens, van Bergen et al. 1998). Our RealTime PCR results suggest that SLPI is expressed in the salivary gland, in the mammary gland, in the lung, testis, spinal chord, bone marrow, colon, kidney and uterus. SLPI expression was significantly higher in the ovarian cancers which led us to believe that the protein levels may be elevated as well. SLPI in mucosal fluids inhibits HIV-1 (Wahl, McNeely et al. 1997) which is why a Dutch company (Hbt HyCult biotechnology, Uden, Netherlands) developed an ELISA to assay SLPI levels in saliva and possibly correlate them with protection against HIV infection. We have used this ELISA kit on serum samples from 10 ovarian cancer patients whose tissues showed high levels of SLPI transcript expression. We paired these results with sera from 10 normal individuals. The assay was performed in the laboratory of Drs Ingegerd and Karl Erik Hellström at the Pacific Northwest Research Institute in Seattle. Figure 8 shows that there is no difference in serum SLPI levels between these two groups.

This disappointing finding correlated well with the fact that SLPI is present in a lot of tissues that could potentially contaminate the blood. The much higher transcript levels in the ovarian tumors do obviously not translate into higher serum levels of the protein.

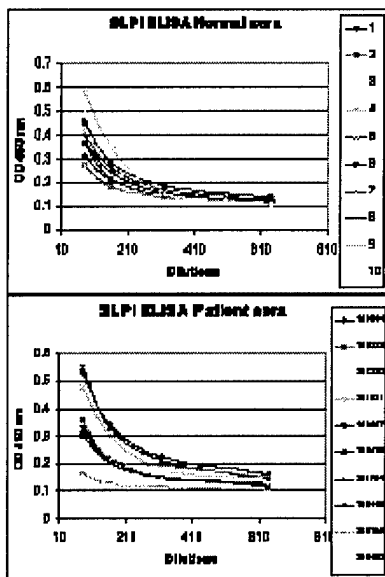


Figure 8 - SLPI ELISA on sera from 10 ovarian cancer patients and 10 controls

Top: sera from 10 normal controls, bottom: sera from 10 ovarian cancer patients whose tissues showed overexpression of SLPI message as assayed by RealTime quantitative PCR (see Figure 7). There is no difference in protein levels between the two groups.

MSLN and WFDC2 serum assays

The laboratory of Drs Hellström at the Pacific Northwest Research Institute have developed an antibody against MSLN used for testing of the presence of MSLN protein in the sera of patients with mesotheliomas (Scholler, Fu et al. 1999). A fusion protein was constructed which was encoded by the WFDC2 gene in combination with genes encoding either a human or mouse immunoglobulin tail. Mice were immunized with the fusion protein having a human tail and hybridomas were made which were screened against the fusion protein with a mouse tail. Monoclonal antibodies were obtained to two different epitopes of the WFDC2 antigen. They were used to construct a double determinant ("sandwich") ELISA, analogous to one for MSLN (Scholler, Fu et al. 1999). Both ELISAs were used to screen 445 sera including 227 healthy controls, 53 ovarian cancer cases of various histologies, and 86 other cancers and benign conditions. All sera were tested for CA125 as control. In depth analysis of the data is under way. Preliminary analysis shows that MSLN, although displaying lower sensitivity than CA125, can add sensitivity to CA125 when used together (Figure 9). This is caused by a small number of sera being positive for either but not both of these 2 markers.

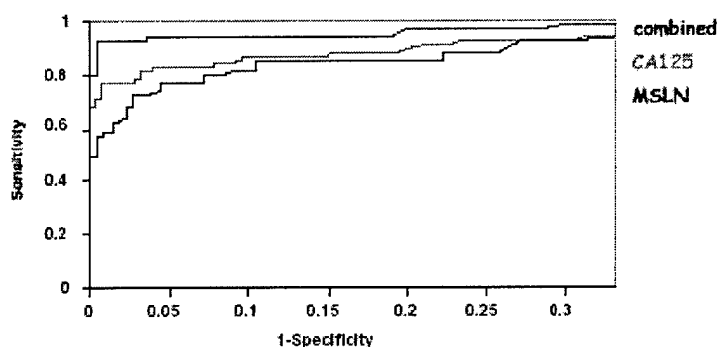


Figure 9 - ROC curve of CA125, MSLN and both markers combined
MSLN adds sensitivity to CA125

WFDC2 results show a low sensitivity of below 50%, but the marker is negative in both healthy controls and in cases with benign ovarian tumors (Table 9). Since both MSLN and CA125 detect benign ovarian tumors next to malignant disease, combining these two markers with WFDC2 increases the specificity of the assay for the detection of malignancy (Figure 10). Currently, ovarian cancer detection is performed by ultrasound (transvaginal sonography, TVS) which picks up all masses, benign and malignant, followed by CA125 which will be negative in some but not all benign cases. The introduction of WFDC2 could potentially identify the truly malignant cases.

Type	Total	HE4 pos%		significant?
healthy controls	277	0	0%	Y
benign ovarian tumors	29	2	7%	N
early stage Ov Ca	19	6	32%	Y
late stage Ov Ca	34	16	47%	Y
all ovarian cancers	53	22	42%	Y
other cancers and benign conditions	86	8	9%	N
TOTAL	445			

Table 9 - Number of sera positive for WFDC2

WFDC2 is negative in both healthy controls and benign ovarian disease cases, making it suitable to add specificity to existing markers.

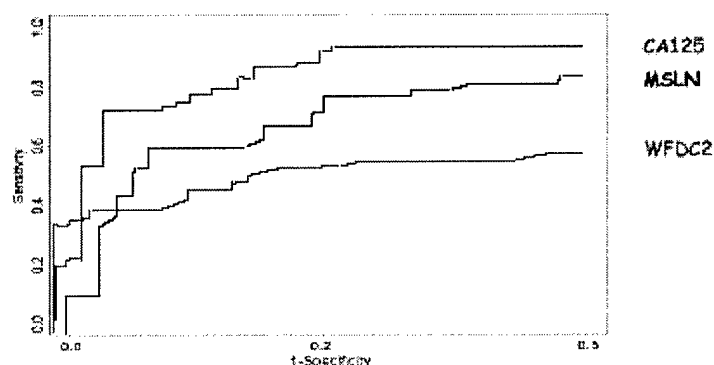


Figure 10 - ROC curve of CA125, MSLN and WFDC2

WFDC2 adds specificity to the combined assay of CA125 and MSLN due to the absence of expression of WFDC2 protein in control sera.

Outlook

We have applied for funding to resequence ~400 clones that either failed to produce a single PCR band or that did not deliver a satisfying sequence. This includes the 7 novel genes which we failed to generate PCR primers for. We have also received funding to generate monoclonal antibodies against additional proteins whose PCR expression profiles suggest that they may be potential marker genes for ovarian cancer: ELF3, EDG7 and GPR39. Within the next month we expect to have one of the two monoclonal antibodies needed for a Sandwich-ELISA. The other antibodies will follow shortly.

Project 1

CONCLUSIONS:

We have identified several potential ovarian cancer marker genes, some of them previously known to be related to cancer (WFDC2, FOLR1, CD24, KRT8, MUC1, LCN2, S100A11, S100A6, IFI27, ERBB2), some with no previous role in cancer (ELF3, GPR39, SLPI, MSLN), some matching to ESTs (22) and some being novel genes (8). These genes and their proteins are currently being evaluated by research laboratories other than ours. The antibody generation is carried out at the Pacific Northwest Research Institute in Seattle by I. and KE Hellström; in situ hybridization on tissue sections and transcript quantitation on cells from peritoneal washes are conducted at Harborview Hospital in Seattle (N. Kiviat). Markers from this study are being employed in a new study funded through the Ovarian SPORC where multiple markers, including CA125, will be evaluated together (Project 4 by Martin McIntosh).

There is little doubt that a useful marker will be a protein that can be found in the blood. It was only during the last 2 years that we discovered the unexpectedly low concordance between mRNA and protein levels of some genes. If we had to repeat the effort of finding a marker again, given the recent advances in proteomics, we would put a lot more emphasis on the protein side. One proposed approach would make use of the ICAT-based protein labeling method to identify membrane-bound proteins in cancer tissues (Gygi, Rist et al. 1999).

REFERENCES:

1. American Cancer Society (1998). American Cancer Society - Cancer Facts and Figures. Atlanta, GA, American Cancer Society Inc.
2. Anderson, L. and J. Seilhamer (1997). "A comparison of selected mRNA and protein abundances in human liver." *Electrophoresis* 18(3-4): 533-7.
3. Ben-Dor, A., L. Bruhn, et al. (2000). "Tissue Classification with Gene Expression Profiles." *The Forth Annual International Conference on Computational Molecular Biology - RECOMB'2000*: 54-64.
4. Ben-Dor, A., L. Bruhn, et al. (2000). "Tissue Classification with Gene Expression Profiles." *J. Comput. Biol.* 7(3): 559-583.
5. Chang, T. J., C. C. Juan, et al. (1998). "Up-regulation of beta-actin, cyclophilin and GAPDH in N1S1 rat hepatoma." *Oncol Rep* 5(2): 469-71.
6. Chen, C. L., S. M. Ip, et al. (2000). "Loss of imprinting of the IGF-II and H19 genes in epithelial ovarian cancer." *Clin Cancer Res* 6(2): 474-9.
7. Davies, J. A., A. D. Perera, et al. (1999). "Mechanisms of epithelial development and neoplasia in the metanephric kidney." *Int J Dev Biol* 43(5 Spec No): 473-8.
8. Desprez, P. Y., D. Poujol, et al. (1992). "Glyceraldehyde-3-phosphate dehydrogenase (GAPDH, E.C. 1.2.1.12.) gene expression in two malignant human mammary epithelial cell lines: BT-20 and MCF-7. Regulation of gene expression by 1,25-dihydroxyvitamin D3 (1,25-(OH)2D3)." *Cancer Lett* 64(3): 219-24.
9. Dressler, G. R. and A. S. Woolf (1999). "Pax2 in development and renal disease." *Int J Dev Biol* 43(5 Spec No): 463-8.
10. Filipek, A., U. Wojda, et al. (1995). "Interaction of calyculin and its cyanogen bromide fragments with annexin II and glyceraldehyde 3-phosphate dehydrogenase." *Int J Biochem Cell Biol* 27(11): 1123-31.
11. Finnegan, M. C., J. R. Goepel, et al. (1993). "Investigation of the expression of housekeeping genes in non-Hodgkin's lymphoma." *Leuk Lymphoma* 10(4-5): 387-93.
12. Fogel, M., J. Friederichs, et al. (1999). "CD24 is a marker for human breast carcinoma." *Cancer Lett* 143(1): 87-94.
13. Geiss, G. K., R. E. Bumgarner, et al. (2000). "Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays." *Virology* 266(1): 8-16.
14. Gygi, S. P., B. Rist, et al. (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." *Nat Biotechnol* 17(10): 994-9.
15. Ichikawa, J. K., A. Norris, et al. (2000). "Interaction of pseudomonas aeruginosa with epithelial cells: identification of differentially regulated genes by expression microarray analysis of human cDNAs." *Proc Natl Acad Sci U S A* 97(17): 9659-64.
16. Kaupilla, S., J. Saarela, et al. (1996). "Expression of mRNAs for type I and type III procollagens in serous ovarian cystadenomas and cystadenocarcinomas." *Am J Pathol* 148(2): 539-48.
17. Kim, J. W., S. J. Kim, et al. (1998). "Increased glyceraldehyde-3-phosphate dehydrogenase gene expression in human cervical cancers." *Gynecol Oncol* 71(2): 266-9.

18. Komatsu, K., A. Andoh, et al. (2000). "Increased expression of S100A6 (Calcyclin), a calcium-binding protein of the S100 family, in human colorectal adenocarcinomas." *Clin Cancer Res* 6(1): 172-7.
19. Lahousen, M., H. Stettner, et al. (1989). "A tumor-marker combination versus second-look surgery in ovarian cancer. I. Clinical experience." *Baillieres Clin Obstet Gynaecol* 3(1): 201-8.
20. Martens, J., J. Baars, et al. (1999). "Can keratin 8 and 17 immunohistochemistry be of diagnostic value in cervical cytology? A feasibility study." *Cancer* 87(2): 87-92.
21. McKee, K. K., C. P. Tan, et al. (1997). "Cloning and characterization of two human G protein-coupled receptor genes (GPR38 and GPR39) related to the growth hormone secretagogue and neurotensin receptors." *Genomics* 46(3): 426-34.
22. Moog-Lutz, C., P. Bouillet, et al. (1995). "Comparative expression of the psoriasin (S100A7) and S100C genes in breast carcinoma and co-localization to human chromosome 1q21-q22." *Int J Cancer* 63(2): 297-303.
23. Naylor, M. S., G. W. Stamp, et al. (1992). "Beta actin expression and organization of actin filaments in colorectal neoplasia." *Epithelial Cell Biol* 1(3): 99-104.
24. Nelson, P. S., W. L. Ng, et al. (1998). "An expressed-sequence-tag database of the human prostate: sequence analysis of 1168 cDNA clones." *Genomics* 47(1): 12-25.
25. Oettgen, P., R. M. Alani, et al. (1997). "Isolation and characterization of a novel epithelium-specific transcription factor, ESE-1, a member of the ets family." *Mol Cell Biol* 17(8): 4419-33.
26. Pedrocchi, M., B. W. Schafer, et al. (1994). "Expression of Ca(2+)-binding proteins of the S100 family in malignant human breast-cancer cell lines and biopsy samples." *Int J Cancer* 57(5): 684-90.
27. Persons, D. A., N. Schek, et al. (1989). "Increased expression of glycolysis-associated genes in oncogene-transformed and growth-accelerated states." *Mol Carcinog* 2(2): 88-94.
28. Pinto, V., M. Marinaccio, et al. (1997). "Preoperative evaluation of ferritinemia in primary epithelial ovarian cancer." *Tumori* 83(6): 927-9.
29. Santala, M., M. Simojoki, et al. (1999). "Type I and III collagen metabolites as predictors of clinical outcome in epithelial ovarian cancer." *Clin Cancer Res* 5(12): 4091-6.
30. Schek, N., B. L. Hall, et al. (1988). "Increased glyceraldehyde-3-phosphate dehydrogenase gene expression in human pancreatic adenocarcinoma." *Cancer Res* 48(22): 6354-9.
31. Scholler, N., N. Fu, et al. (1999). "Soluble member(s) of the mesothelin/megakaryocyte potentiating factor family are detectable in sera from patients with ovarian carcinoma." *Proc Natl Acad Sci U S A* 96(20): 11531-6.
32. Schummer, M., W. V. Ng, et al. (1999). "Comparative hybridization of an array of 21 500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas." *Gene* 238: 375-385.
33. Tanaka, M., K. Adzuma, et al. (1995). "Human calgizzarin; one colorectal cancer-related gene selected by a large scale random cDNA sequencing and northern blot analysis." *Cancer Lett* 89(2): 195-200.
34. Toffoli, G., C. Cernigoi, et al. (1997). "Overexpression of folate binding protein in ovarian cancers." *Int J Cancer* 74(2): 193-8.
35. Tokunaga, K., Y. Nakamura, et al. (1987). "Enhanced expression of a glyceraldehyde-3-phosphate dehydrogenase gene in human lung cancers." *Cancer Res* 47(21): 5616-9.

36. Tripathi, P. K. and S. K. Chatterjee (1996). "Elevated expression of ferritin H-chain mRNA in metastatic ovarian tumor." *Cancer Invest* 14(6): 518-26.
37. Tymms, M. J., A. Y. Ng, et al. (1997). "A novel epithelial-expressed ETS gene, ELF3: human and murine cDNA sequences, murine genomic organization, human mapping to 1q32.2 and expression in tissues and cancer." *Oncogene* 15(20): 2449-62.
38. Urban, N. (1999). "Screening for ovarian cancer. We now need a definitive randomised trial." *Bmj* 319(7221): 1317-8.
39. Van Ginkel, P. R., R. L. Gee, et al. (1998). "The identification and differential expression of calcium-binding proteins associated with ocular melanoma." *Biochim Biophys Acta* 1448(2): 290-7.
40. Wahl, S. M., T. B. McNeely, et al. (1997). "Secretory leukocyte protease inhibitor (SLPI) in mucosal fluids inhibits HIV-I." *Oral Dis* 3 Suppl 1: S64-9.
41. Westin, U., A. Polling, et al. (1999). "Identification of SLPI (secretory leukocyte protease inhibitor) in human mast cells using immunohistochemistry and in situ hybridisation." *Biol Chem* 380(4): 489-93.
42. Wingens, M., B. H. van Bergen, et al. (1998). "Induction of SLPI (ALP/HUSI-I) in epidermal keratinocytes." *J Invest Dermatol* 111(6): 996-1002.
43. Yin, D. L., L. Pu, et al. (1998). "Antisense oligonucleotide to insulin-like growth factor II induces apoptosis in human ovarian cancer AO cell line." *Cell Res* 8(2): 159-65.
44. Yuan, C. C., H. T. Ng, et al. (1988). "Hyperferritinemia in ovarian cancer." *J Reprod Med* 33(2): 193-5.
45. Yun, K., M. Fukumoto, et al. (1996). "Monoallelic expression of the insulin-like growth factor-2 gene in ovarian cancer." *Am J Pathol* 148(4): 1081-7.

Project 2

Antibody Immunity to Cancer Related Proteins as a Serologic Marker for Ovarian Cancer

Nicole Urban, ScD, Brad Nelson, Ph.D., Mary L. Disis, MD

INTRODUCTION

Early diagnosis is essential to make progress in the treatment of and, ultimately, survival from ovarian cancer. Serologic markers, such as CA-125, can potentially indicate the presence of ovarian cancer. However, like many serum markers, CA-125 is shed from the surface of growing tumor and, in general, is associated with bulky disease. A serologic marker that is prevalent and readily detected in early-stage disease would be a more optimal candidate to develop as a screening tool.

The immune system has evolved to detect proteins that are abnormal in terms of primary sequence, overexpression, tissue context or inflammatory context. A large number of tumor proteins are abnormal by these criteria and hence trigger T cell and B cell responses in cancer patients. Examples of tumor antigens that are common to a number of different cancers, including ovarian cancer, are p53 and HER2/neu. Studies in breast cancer have shown that tumor-specific antibody responses to p53 and HER2/neu can occur early during tumorigenesis. Moreover, tumor-specific antibodies can be detected by simple and inexpensive ELISA-based blood tests. For these reasons, we are investigating whether serum antibody responses to ovarian tumor antigens could potentially serve as indicators of early-stage disease.

We hypothesize that women with ovarian cancer will demonstrate serum antibody responses to one or more ovarian tumor antigens, and that such responses will be rare or absent in women with benign ovarian disease and normal controls. This hypothesis was tested by first analyzing antibody responses to two known tumor antigens (p53 and HER2/neu) in women with malignant and benign ovarian disease and normal controls. Second, we used an immunoscreening technique known as SEREX to discover new tumor antigens that are recognized by serum antibodies in women with ovarian cancer. We are currently assessing the prevalence of antibody responses to new antigens among cases and controls, as well as the extent of overlap with responses to p53 and HER2/neu. The long-term goal continues to be to assemble a panel of ovarian tumor antigens that constitute a sensitive and specific blood test for early-stage ovarian cancer.

BODY

Task 1: Perform ELISA screens for p53, HER2/neu and Myc **Months 1-24:**

A. An ELISA based screen will be used to probe serum from ovarian cancer patients and control individuals for the presence of antibodies against the tumor associated proteins p53, H2N and Myc. It is anticipated to perform this set of tests on 350 cases per year.

1. To develop reproducible assays for detecting HER2 antibodies. Data presented in the first year's report demonstrated that near CLIA grade assays have been developed for the detection of HER2 specific antibodies based on a capture ELISA format. *****Table 1 demonstrates long term validation data on both the HER2 and p53 antibody assays. Calculations were made on over 100 plates analyzed over 14 months. As previously reported, peptide assays and recombinant protein assays did not prove superior to the capture ELISA format developed. All blood samples collected through the ORCHID study have been analyzed for HER2 antibodies. In addition, a reference population of 175 volunteer blood donors has been analyzed. Results of the final analysis are described below (Task 5).

2. To develop reproducible assays for detecting p53 antibodies. Data presented in the first year's report demonstrated that near CLIA grade assays have been developed for the detection of p53 specific antibodies based on a capture ELISA format. *****Table 1 demonstrates long term validation data on both the HER2 and p53 antibody assays. Calculations were made on over 100 plates analyzed over 14 months.

In addition, we synthesized the 2 putative immunodominant B cell epitopes of p53 (see previous report). An indirect ELISA was developed. In 96-well microtiter plates (Dynex Technologies, Inc., Chantilly, VA), columns were coated with the p53 peptide, at a concentration of 20 µg/ml, diluted with carbonate buffer and added at 50 µl per well. Alternating columns were coated with 50 µl/well of carbonate buffer alone. The standard curve column, column 12, was incubated with the purified IgG titrations as above, at 40C overnight. After overnight incubation, all wells were blocked with 1% casein/PBS, 100 µl/well and incubated at room temperature on a rocker for 1-2 hours. Plates were then washed with a 0.15%casein/1% PBS/0.05% Tween-20 wash buffer 4 times before serum diluted in 10%FCS/PBS/1% BSA/25µg/ml mouse IgG at 1:100, 1:200, 1: 400 and 1:800 dilutions. Plates were incubated for 2 hours at room temperature on a rocker. Plates were then washed 4 times with casein-based wash and incubated for 45 minutes at room temperature on a rocker after addition of 50 µl/well IgG-HRP conjugate diluted 1:10,000 in PBS/BSA buffer. After a final 4 washes with casein-based wash buffer, TMB reagent was added 75 µl/well and color reaction read at 640nm until the well containing the 0.16 µg/ml standard reached an OD of 0.3. Reaction was then stopped with 75 µl/well 1N HCL and read at 450nm. The OD of each serum dilution was calculated as the OD of the peptide-coated wells minus the OD of the buffer-coated wells. Values for delta OD were calculated from the log-log equation of the line for the standard curve on each plate. Samples that returned a positive delta OD for 3 of 4

dilutions were counted, and a positive sample was defined as a $\mu\text{g/ml}$ value greater than the mean of the normal population plus 3 sd.

ASSAY VALIDATION

- Normal Range: 50 serum samples from normal donors were assayed by peptide ELISA and a normal range established by determining the mean and standard deviation of all samples and calculating a cut-off value of the mean plus 3 standard deviations, a confidence interval of approximately 99%. The peptide ELISA returned a normal mean and standard deviation of $0.052 \pm 0.11 \mu\text{g/ml}$, giving us a cut-off value of $0.382 \mu\text{g/ml}$. We found that 1% of our normal samples resulted positive for any peptide.
- Accuracy: The peptide ELISA returned an average CV of 11%.
- Precision: The peptide ELISA returned an intra-assay precision and interassay precision of 9% and 17%, respectively.

Samples from 40 breast cancer patients (archived) were analyzed for p53 protein by the standard assay and the by the p53 peptide assays as described. 20% of the patients had antibody responses to p53 using the capture ELISA method. 13% had antibodies to one or both of the peptides. 2 of those patients did not have detectable p53 antibodies by protein assay. 3/5 of the peptide specific responses could be validated by Western blot. Therefore, we determined that the p53 antibody assay in the capture ELISA format (Table 1) was a more robust determination of pre-existent antibody immunity to p53. Further studies on the peptides will be undertaken as larger populations of p53 antibody positive patients are identified in subsequent studies.

All blood samples collected through the ORCHID study have been analyzed for p53 antibodies using the capture ELISA format. In addition, a reference population of 175 volunteer blood donors has been analyzed. Results of the final analysis are described below (Task 5).

Table 1

PARAMETERS EVALUATED	RESEARCH ASSAYS		CLIA-BASED STANDARD
	HER2	P53	
Accuracy	12%	10%	<10%
Precision			
Intra assay	9%	12%	<10%
Inter assay	20%	15%	<10%

Specificity	77%	100%	>80%
Sensitivity	89%	93%	>90%
Linearity	r=0.98	r=0.95	r=0.95

3. To develop reproducible assays for detecting c-myc antibodies. Antibodies against c-myc have been reported in patients with cancer. Critical to the development of antibody assays detecting c-myc is the ability to validate responses by western blot or the availability of monoclonal antibodies to use as coating antibodies in a capture ELISA. To date we have not found an antibody to c-myc that will work reproducibly in ELISA. Likewise, the specimen core has had difficulty analyzing specimens for c-myc expression. Therefore, development of a c-myc antibody assay was abandoned. Work in the last 6 months has instead focused on the assessment of CA-125 as a potential antibody target. Preliminary studies using commercial antibodies and ovarian cancer cell lines expressing the glycoprotein demonstrates the capture ELISA template is feasible and Western blot analysis can be reproducibly performed. Experiments evaluating a reference population of volunteer blood donors are underway.

Task 2: Determine SEREX baseline
Months 1-6:

A. Conduct ten serial absorptions on sera from three normal individuals and three ovarian cancer patients with known reactivities to one or more of the p53, H2N and Myc antigens.

The goal of this task was to optimize the signal-to-noise ratio of the SEREX protocol and to pre-clear serum samples of antibodies to E. coli. Experiments performed in Months 1-6 led to a reliable pre-clearing procedure. Serum samples are first incubated overnight with matrix-immobilized protein lysates from E. coli. Serum is further pre-cleared by two serial incubations with nitrocellulose membranes containing empty lambda phage on a lawn of E. coli. This pre-clearing procedure has successfully reduced background reactivity to acceptable levels in over 80 serum samples from cancer patients and normal controls and has become our laboratory standard.

B. Construct a cDNA expression library from pooled ovarian tumor samples.

Ten stage III/IV serous ovarian tumors were used to construct a cDNA library using the lambda phage vector lambda TriplEx. Library construction proceeded as planned. In addition, in Year 3 a second cDNA library was constructed in the vector lambda Zap II XR using RNA from one human testes. The rationale for constructing this latter library was that testes tissue expresses an extremely broad range of human mRNAs and therefore serves as an excellent source of rare mRNAs. Indeed, many of the most promising tumor antigens discovered to date are expressed also in testes, but not other normal tissues. Such gene

products are now referred to as cancer-testes antigens, or "CT antigens". Examples include many members of the MAGE superfamily. We reasoned that by SEREX screening a testes cDNA library with serum from ovarian cancer patients, we might identify CT antigens and other rare antigens that may be useful for early detection.

C. Assess the quality of the library.

Titration experiments demonstrated that both the ovarian tumor and testes cDNA libraries contain $>1 \times 10^6$ primary clones. PCR analysis demonstrated $>95\%$ recombinant phage and an average insert size of 1.5 to 1.7 kb. Partial sequencing of randomly picked clones showed no evidence of genomic or bacterial DNA contamination. Most important, both libraries have been used successfully for SEREX screening with serum from over 30 ovarian cancer patients (discussed below).

Task 3: Use SEREX to screen serum from ovarian cancer patients **Months 6-20:**

A. Identify novel ovarian tumor antigens.

Serum samples from 29 ovarian cancer patients were used for primary SEREX screening of the ovarian tumor cDNA library. The library was plated on a lawn of Y1090⁻ E. coli cells at a density of 2.5×10^3 pfu per 100 mm NZYCM plate (NZ amine + yeast extract + 1% casamino acids + 2% MgSO₄ + 1.5% agar; Sigma). When phage plaques first became visible (~3h), plates were overlaid with IPTG-impregnated nitrocellulose membranes and incubated at 37°C. After overnight growth, the membranes were removed, washed in TBST (Tris buffered saline [TBS] + 0.05% Tween-20), and blocked in TBS + 1% BSA (bovine serum albumin). Membranes were then incubated overnight at room temperature with pre-cleared patient serum diluted 1:100 in TBS + 1% BSA. The following day, membranes were washed with TBS, and incubated for 45 minutes with alkaline phosphatase-conjugated goat anti-human antibody (specific for IgG, IgA and IgM; Pierce) diluted 1:7500 in TBS + 1% BSA. Membranes were developed with NBT/BCIP. As shown in ****Fig. 1, positive phage plaques typically appeared as darkened halos or spots. All presumptive positive phage from the primary screen were picked from the original agar plates and stored at 4°C in SM buffer (150 mM NaCl + 10 mM magnesium).

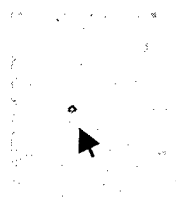


Fig 1. Example of an immunoreactive phage plaque from a primary SEREX screen. The arrowhead indicates a single immunoreactive plaque (dark halo) amongst several hundred non-reactive plaques (clear spots).

To date, we have performed primary SEREX screening on approximately 4.7×10^6 phage plaques of the ovarian tumor library with serum from 29 ovarian cancer patients. In addition, we have screened the recently constructed testes cDNA library with serum from 13 ovarian cancer patients. On average, 1-2 immunoreactive phage are found for every 2,500 phage screened. Upon re-screening with secondary antibody alone, the majority (~90%) of positive phage from the ovarian tumor library were found to encode IgG molecules, which presumably are derived from B cells in the original tumor samples. To rapidly exclude these unwanted phage clones and identify those rare clones with cancer-specific immunoreactivity, we developed a novel array-based method to perform rapid, reproducible, and well-controlled secondary SEREX screens. Two-dimensional arrays containing up to 100 phage plaques were constructed by placing 1 μ l drops of phage suspension in a grid-like pattern onto a lawn of *E. coli* on a rectangular agar plate. As with primary SEREX screens, plates were then overlaid with IPTG-impregnated nitrocellulose membranes and incubated at 37°C overnight. Under these conditions, the vast majority of phage give rise to a single plaque approximately 0.5 cm in diameter, irrespective of the exact titer of the original phage solution. Each membrane thus contains up to 100 individual phage plaques, each with a defined position on a grid. Negative and positive control phage are included for comparison with test phage. Multiple replicates of these arrays can be rapidly constructed using a multi-channel pipettor. Nitrocellulose membranes containing such phage arrays were then immunoblotted with serum samples using the same methodology as for primary SEREX screens (above).

Fig. 2 shows a subset of the array results obtained from our ovarian cancer study. The figure shows two replicate arrays containing 42 phage plaques. A non-recombinant (i.e., empty) phage was plated at several positions to serve as a negative control. The leftmost array was immunoblotted with serum from an ovarian cancer patient, whereas the rightmost array was immunoblotted with serum from a normal control (a female over the age of 30 with no personal history of cancer). The arrowheads show 6 phage that were reactive with serum antibodies from the cancer patient, but not the normal control. Anywhere from 1-5% of phage from the primary screen show cancer-specific immunoreactivity such as this across the panel of case and control sera. The 6 phage clones in Fig. 2 were subjected to standard DNA sequencing. As summarized below, 4/6 were found to encode the tumor suppressor p53, and 2/6 encoded a novel zinc finger-containing protein called hZF5.

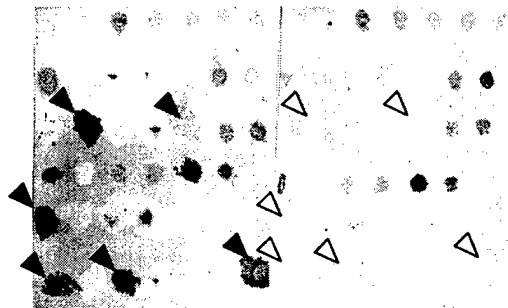


Fig 2. Secondary SEREX screening of ovarian cancer cDNA clones by phage array. The left and right panels show duplicate nitrocellulose membranes containing a 2-D array of recombinant phage clones that were identified in a primary SEREX screen of an ovarian tumor cDNA library. The left panel was immunoblotted with

serum from a ovarian cancer patient (stage III , serous) whereas the right panel was immunoblotted with serum from a normal control. Membranes were then probed with a human IgG-specific, AP-conjugated secondary antibody and developed with NBT/BCIP. Immunoreactive phage plaques appear as dark circles, whereas non-reactive phage are clear. The arrows indicate 6 phage that showed a cancer-specific pattern of immunoreactivity with these and other serum samples.

To date, our SEREX screening efforts involving both the ovarian and testes cDNA libraries have identified 26 different gene products that appear to have a cancer-specific pattern of immunoreactivity (when tested in a SEREX array format using sera from 30 cancer cases and 20 normal controls, as described below). The 26 antigens are summarized in Table 2. Among these antigens are the tumor suppressor p53 and the cancer-testes antigen NY-ESO-1, both of which are well-documented tumor antigens. None of the other antigens has previously been shown to be immunogenic in cancer. Among these is a protein called Ubiquilin-1 that has homology to Ubiquitin but has no ascribed function. Intriguingly, one of the antigens we identified by SEREX (IFI27) was also identified by HDAH in Project 1 (see Table 7 in the Project 1 report) due to its overexpression at the mRNA level in ovarian cancer. This suggests that the immunogenicity of IFI27 might be attributable to overexpression by ovarian tumors, leading to broken peripheral tolerance to this self protein.

Table 2. List of antigens discovered by SEREX immunoscreening, and frequency of antibody responses to these antigens among patients with ovarian cancer and normal controls.

Antigen	Ovarian cases (n=54)	Controls (n=20)
p53	6	0
TOP2a	3	0
RUVBL	2	0
KIAA0035	1	0
HCAP-G	1	0
DLD	1	0
DDX9	1	0
STMN1	1	0
ILF3	1	0
NY-ESO-1	10	0
UBQLN1	3	0
HOXB6	3	0
ZFP161	1	0
HIS1	2	0
SPARC	1	0
CD44	1	0
YB-1	1	0
FBXO21	1	0
FLJ20267	2	0
DDX5	2	0
FLJ22318	1	0
KNSL6	2	0
FLJ10534	1	0
NKTR	1	0
IFI27	2	0
HSP40	1	0

Current screening efforts are focused on patients who are negative for antibody responses to the 26 antigens identified so far. At this time, we do not know whether these patients are completely deficient in antibody responses to antigens represented in the libraries, or whether continued screening will reveal novel antigens to which they respond. The longterm goal is to identify additional antigens that allow detection of this patient subset, so as to increase the overall sensitivity of the antigen panel for detecting ovarian cancer.

B. Prioritize the evaluation of novel ovarian tumor antigens. Statistical methods will be applied to identified antigens to determine if the discovery looks promising for translation into a test for use in the general public.

As described above, we are using SEREX-based arrays for initial prioritization of ovarian tumor antigens discovered through SEREX. All 26 antigens discovered to date have been arrayed and exposed to serum from 54 ovarian cancer patients and 20 normal controls. Table 3 shows results for a subset of these patients (n=28), scored against the most frequently recognized antigens. It can be seen that some patients showed a response to only one antigen in the panel, whereas others showed responses to several antigens. Likewise, some antigens

were recognized by multiple patients, whereas others were recognized by only a single patient. Importantly, 10/28 (36%) of patients showed an antibody response to at least one antigen in the panel. Based on these results, we have prioritized NY-ESO-1, Ubiquilin-1, HOXB6 and TOP2a for follow-up ELISA studies with larger numbers of case and control sera.

Serum samples

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	Total
NY-ESO-1																													5
HOX B6																													2
HIS 1																													1
UBQLN1																													1
p53																													4
TOP2a																													1
KNSL6																													1
FLJ20627																													1
Total:	0	1	1	0	0	2	1	1	0	3	0	0	0	0	0	0	3	0	1	0	0	0	1	0	0	0	0	2	

Table 3. Pattern of IgG serum antibody responses to a subset of SEREX-defined antigens among 28 late-stage serous ovarian cancer patients. Each column shows the results for one ovarian cancer patient (numbered 1-28). Black cells indicate a positive response, as detected by SEREX-based arrays. All antigens were negative when tested against a panel of 20 normal control sera from age-matched women

Task 4: Perform ELISA screens for promising candidates

Months 18-24:

A. An ELISA based screen will be used to probe serum for the presence of antibodies against promising candidates that are identified by SEREX technology in Project 2.

We assembled a full-length cDNA clone encoding NY-ESO-1 and produced His-tagged recombinant protein in the mammalian cell line COS-7. The cDNA was inserted into the mammalian expression vector pcDNA4.1/HisMAX (Invitrogen), which fused six histidine residues to the N-terminus of the protein. The resulting plasmid was transiently transfected into COS7 cells using lipofectamine-Plus. *****Fig. 3 shows an anti-His Western blot of COS-7 lysates containing recombinant histagged NY-ESO-1 and, as controls, histagged p53 and LacZ.

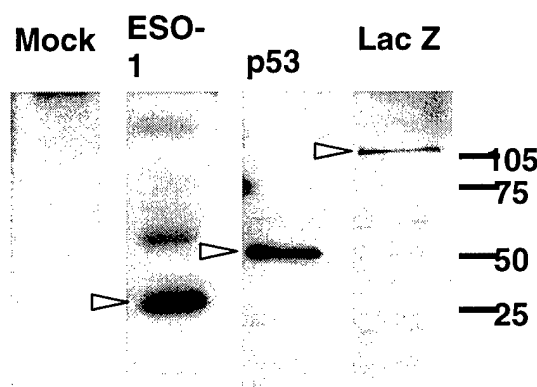


Fig 3. Western blot showing expressing of His-tagged recombinant tumor antigens in mammalian COS7 cells. Cells were transiently transfected with pcDNA3-based expression vectors encoding His-tagged ESO-1, p53 or Lac Z (as a control). Mock transfected cells served as a negative control. Nuclear extracts were prepared, subjected to SDS-PAGE and immunoblotted with a monoclonal antibody to the His tag (Sigma). Antibody detection was by enhanced chemiluminescence. Recombinant proteins are indicated by open arrowheads.

These COS7 cell lysates were used as antigen sources in ELISA to assess serum antibody responses to p53 and NY-ESO-1 in a subset of ovarian cancer patients. Briefly, 96-well nickel-coated ELISA plates (Clontech) were blocked with PBS/1% BSA, washed with PBS/0.5% Tween-20 and then be incubated with lysates (10^8 cells/20 ml PBS) from COS7 cells transfected with plasmids encoding NY-ESO-1, p53 or LacZ, or empty pcDNA4.1/HisMAX to serve as a negative control. After washing, plates were incubated with serum at 1:50 in PBS/1% BSA. After washing, plates were incubated with goat anti-human antibody conjugated to horseradish peroxidase (HRP). Plates were developed with TMB and read at 450 nm. As shown in Fig. 4, in this preliminary experiment, four patients showed antibody responses to NY-ESO-1, and three showed responses to p53. This ELISA method will be used in future to assess the serum antibody response to NY-ESO-1 in larger groups of ovarian cancer patients and controls, as per Aim 1.

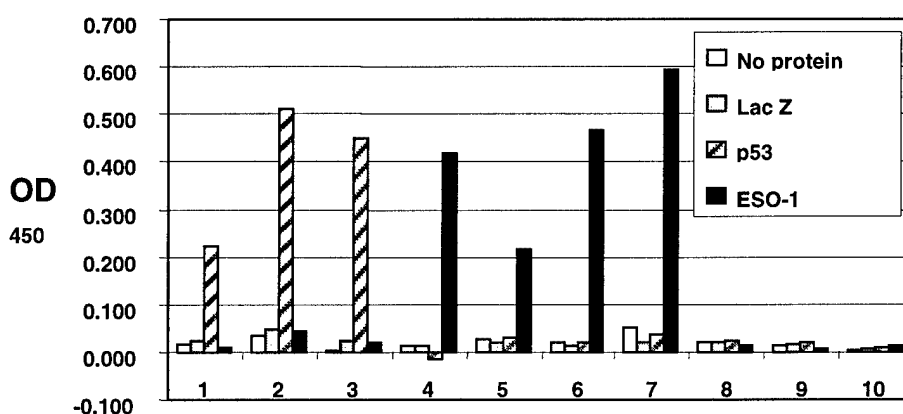


Fig 4. ELISA demonstrating serum antibody responses to p53 and ESO-1 in patients with ovarian cancer. Lysates from COS7 cells (see Fig. 3) expressing His-tagged p53, ESO-1 or, as a negative control, Lac Z were added to nickel-coated ELISA plates. After unbound proteins were washed away, serum from 10 ovarian cancer patients was added at 1:50 dilution, followed by HRP-conjugated goat anti-human IgG secondary antibody.

Plates were developed with TMB and read at 450 nm. Patients #1-3 show a serum antibody response to p53, whereas patients #4-7 show a response to ESO-1. Patients #8-10 show no response to either protein.

In addition to NY- ESO-1, we have recently produced recombinant Histidine-tagged versions of Ubiquilin-1, p53, KNSL6, HIS1 and HOXB6. Some of these antigens did not express well in COS-7 cells and therefore a prokaryotic expression system was used instead. We have also collected full-length cDNA for each of the antigens listed in Figure 2, and anticipate the development of HIS-tagged expression constructs and ELISAs for all of these antigens. Once adequate ELISA conditions are established for assessing serum antibody responses to these antigens, we will commence large-scale ELISA experiments as per Aim 1.

B. Full length cDNA will be obtained and an ELISA based screen will be developed for at least one of the most promising overexpressed genes discovered through HDAH in Project 1. This task will not be completed by month 24.

Studies are planned to assess the serum antibody response to HE4, which was discovered in Project 1. This work will commence when specific antiserum to HE4 becomes available for use in sandwich ELISA.

Task 5: Pool data for analysis

Months 6-24:

A. All discovery data will be combined with data from the other labs through the coordination core. The Project Director will summarize on a routine basis the results and provide them to the Coordination Core for further interpretation and incorporation.

Please refer to page 66 for data analyses conducted by the Statistical, Clinical and Laboratory Coordinating Core.

CONCLUSIONS

In addition to the known tumor antigens p53 and HER2/neu that were evaluated in Aim 1, a large number of novel candidate antigens that are immunogenic in ovarian cancer have been identified by SEREX. We are now poised to evaluate serum antibody responses to the combined panel of antigens using large numbers of sera from patients with malignant and benign ovarian disease and normal controls. Our preliminary results using SEREX-based arrays indicate that approximately 36% of ovarian cancer patients show an antibody response to at least one antigen in our panel. While 36% represents a lower proportion of patients than can be detected by CA125, the specificity of this test appears to be very high (0/20 normals scored positive). Moreover, continued SEREX screening is expected to provide additional antigens that may increase the overall sensitivity of the panel. Future studies will assess whether this screening strategy is complementary to existing strategies such as CA125 and transvaginal ultrasound for the detection of early-stage ovarian cancer.

In addition to their potential utility for early detection of ovarian cancer, at least one of these antigens (NY-ESO-1) shows promise as a target for immunotherapy, therefore we have also launched efforts toward this goal. Finally, related studies of breast and colorectal cancer have received funding and been initiated as a result of this work on ovarian cancer.

APPENDICES

1. Copy of manuscript:

Stone, B., Schummer, M., Paley, P.J., Crawford, M., Ford, M., Urban, N. and Nelson, B.H. 2001. MAGE-F1, a novel ubiquitously expressed member of the MAGE superfamily. *Gene* 267:173-182.

2. Copy of provisional patent:

U.S. Provisional Patent Application Nos. 60/278,253 and 60/278,237. Tumor Associated Antigens and Methods for Using the Same.

Statistical, Clinical and Laboratory Coordinating Core

Nicole Urban, ScD, Garnet Anderson, Ph.D., Nancy Kiviat, MD, Leona Holmberg, MD, Jane Kuypers, Ph.D., Charles Drescher, MD, Mary Anne Rossing, Ph.D.

INTRODUCTION

The purpose of the Core shared resource was to support the work of the project investigators by collecting, storing and providing tissue and blood for analyses, as well as conduct of statistical analysis of project results. The specific aims of the Statistical, Clinical and Laboratory Coordination Core were:

- To develop a resource of well-characterized women with associated data, blood and tissue specimens that will be used to develop and test new markers for disease as specified in Projects 1 and 2.
- To characterize the blood and tissue from these women with respect to CA-125 levels; expression of p53, HER2/neu, and Myc; and histology.
- To provide statistical design and analysis support for Projects 1 and 2.
- To describe the joint behavior of novel and previously established markers, and to investigate the relationship of these markers to other clinical factors (e.g., stage at diagnosis, history of cancer, sonography findings) and demographic or epidemiological data (age, menopausal status, number of ovulatory cycles).

The development of this resource of data and specimens has been essential to all phases of the work conducted by the Projects, and laid the foundation for Core Shared Resources in our ovarian cancer Specialized Program of Research Excellence. The Core assured participant data and specimens were collected in a standardized manner, with proper attention paid to the collection protocols to assure the quality of the specimens for the proposed assays. Standardization of specimen collection ensured that all of the specimens analyzed in Projects 1 and 2 were sufficiently and uniformly characterized to allow for the reliable and valid interpretation of project results. In addition, by overlaying the statistical design for the assays of both projects and obtaining the data from the Core laboratory on the same population, we were able study the inter-relationship of both newly identified and established markers. The tasks and status of each task proposed in the original statement of work is included as Appendix A.

Identification and Recruitment of Participants

All patient recruitment for the study occurred at the office of Pacific Gynecology Specialist (PGS), located on the campus of Swedish Medical Center (SMC) in Seattle. Approximately 60% of the PGS oncology practice occurs on the campus of SMC, while the remainder is divided among seven community hospitals in the Puget Sound area. The recruitment and

specimen collection staff for this study were housed at the Marsha Rivkin Center for Ovarian Cancer Research, providing close proximity to and interaction with the clinical community and patient recruitment being conducted for the study.

All potential participants were invited to participate in the study by the attending physician or study nurse at the time of their pre-operative office visit. Interested patients were provided for their review a brochure describing the study, known by its acronym ORCHID – Ovarian Research Collaboration Helping to Improve Detection. The brochure is included as Appendix H. If the patient agreed to participate in the study, the attending physician, nurse or a research study staff member reviewed the consent form and other required enrollment documents with the patient.

Completed enrollment forms were returned to the Marsha Rivkin Center, at which time data from enrollment forms was entered into the study database, with each participant assigned a unique participant number (UPN).

An additional group of women undergoing surgery at other hospitals where tissue collection protocols had not yet been developed were recruited to provide blood during the pre-operative office visit (“blood only collections”). Collection of blood from this group of women began in 1999 and while not as successful as the tissue component of the study, the piloted operational procedures for blood only collections have been honed further in the ovarian SPORE grant. As a result of the DOD funded study experience, SPORE investigators have developed and expanded blood only recruitment protocols to enroll post-chemotherapy patients in serial blood collections, to enroll healthy controls for blood donation, and to enroll high-risk women for screening. These protocols are included as Appendix M .

Data Collection, Warehousing and Storage

Development and use of Data Collection Instruments

For this study, data collection instruments were created to ensure that information needed for scientific research as well as that required for human subjects participation was obtained. The data collection instruments utilized by the Coordinating Core can be classified into three basic categories: Enrollment, Specimen Collection and Clinical Data Requirements.

At the time of enrollment, the participant completed a combined enrollment/medical records release form, an informed consent form and a self-administered 20-minute questionnaire. The self-administered questionnaire completed by all study participants was developed by the study epidemiologist, Mary Anne Rossing, Ph.D. The data collected focus on known or suspected epidemiological risk factors for ovarian cancer, including: menstrual and reproductive history; use of exogenous hormones (oral contraceptives and hormone replacement therapy); family and personal history of ovarian, breast and other cancers; sociodemographic factors; history of selected gynecologic surgeries including hysterectomy and tubal ligation; and other relevant medical conditions. The data collected was sufficient to categorize women according to their menopausal status, estimated total years of ovulation,

and prior personal and family history of cancer. Data also allowed for future examination of the possible relation of various medical, hormonal, and reproductive factors with levels of various cancer screening markers of interest. Of the 346 participants enrolled in this study, 303 (87%) fully completed the study questionnaire.

Data collection instruments were also created for all specimen collection, requisition, and characterization. These forms captured data entered into the specimen tracking system, including specimen type, site, processing method and location. A unique 6-digit number was used to label all specimens, with a duplicate of the label attached to the specimen collection form, blood processing form and specimen tracking form. These forms are labeled as Appendix I in the Appendices included with this report. The unique label number for each specimen is linked to each participant's UPN in the specimen inventory database.

Extensive characterization was conducted for all specimens collected for this study. Clinical data collection forms were created to classify the specimens stored in the study repository. The study automatically received a pathology report and operative notes on patients who provide specimens. Dr. Charles Drescher conducted the first level of review by assigning a clinical diagnosis to the patient based on the pathology report and operative notes. A second level review, specific to the sites sampled, was conducted by Dr. Nancy Kiviat at the Core facility whereby all formalin-fixed, paraffin-embedded specimens which correspond directly with the fresh frozen tissues were examined for histology.

In addition to histologic classification, the tissue was examined utilizing Immunohistochemistry techniques which is described in detail later in this report. A chart review was also conducted whereby relevant clinical data was abstracted and entered into the database system. The Data Coordinator reviewed the chart to abstract additional data regarding probable diagnosis prior to surgery; type and date of diagnostic tests performed and test results. All of the data captured above was entered into the tracking system and associated at the participant or specimen level. Examples of these forms are included as Appendix J in the Appendices.

Development of the Data Management and Tracking Systems

As part of this study, we developed a relational database management system programmed in Visual FoxPro and Visual Basic to support all the tracking functions and store all the key scientific data associated with this project. This information system consists of two key components: 1) a participant enrollment database, and 2) a specimen inventory and data tracking system. Enrollment data, including personal contact information, is stored in the participant enrollment database. At the time of enrollment entry, the database generated a unique participant number (UPN) that is used in all subsequent correspondence regarding the study participant, blood and tissue specimens, and the data associated with these specimens.

The specimen tracking system is a multi-functional application that tracks all specimens collected for the QUEST study as well as for ORCHID and the ovarian SPORE program. Each individual specimen container or vial is labeled with a unique 6-digit number at the time

of collection; these numbers are in turn associated with the UPN of the donating participant and the date of collection at the time of data entry.

Immunohistochemistry data and histology characterizations are also stored in this component. The specimen tracking system also allows for the entry of clinical data relevant to the distribution of tissue and blood specimens, including surgical pathology diagnoses, pre-operative CA 125 results, and other relevant clinical conditions. This system also serves as an inventory database allowing us to track the location of specimens in the freezers and those that have been sent to project investigators.

The original specimen tracking system platform has since been expanded under the auspices of the ovarian SPORE program. Routine reports can now be generated to monitor specimen accrual, histologic subtype and freezer location. An algorithm to automate creation of “pull lists” for specimen distribution was developed and has greatly increased efficiency and accuracy in pulling large quantities of specimens for analyses. Examples are included as Appendix Items K and L.

Both components of the data management system reside on password-protected servers managed by the IS staff of the Cancer Prevention Research Program at Fred Hutchinson Cancer Research Center. In addition to the logging onto the network, a staff member entering the ORCHID system must supply a unique password to run the application.

Collection of epidemiologic and clinical data

Epidemiologic and health history data was routinely collected from all participants. Each week, study staff generated a report showing successful collections without other required data (questionnaires, core histologic review, clinical data records etc). For self-administered forms, the project coordinator follows procedures outlined in the Follow Up Protocol. For clinical information, a clinical data follow up reported was created by the study database. A regular chart review of recently enrolled participants was conducted by study staff to obtain detailed information on final diagnosis. A copy of the final pathology report was automatically obtained from Dynacare Laboratory of Pathology and included in the participant's study file. Data from the Core Laboratory on specimen characterization was submitted bi-weekly in electronic format for inclusion in the database.

As described previously, clinical follow-up data is collected via review of the participant's medical records using standardized forms developed by the Core investigators, and entered into the clinical database by study staff. This data included selected information on disease characteristics including diagnostic test results, histology, stage, grade, tumor distribution, extent of residual disease and any other standardized data as determined by Core investigators. Procedures for clinical follow up data extraction piloted in the DOD funded study has been expanded in the ovarian SPORE.

During the last year, we received Institutional Review Board approval to retrospectively characterize chemotherapy response on all participants who donated specimens under the

ORCHID protocol. The information collected included semi-annual follow-up of participants regarding chemotherapy administered, response to treatment, disease status and survival. The information abstracted was sufficient to characterize specimens for use in a SPORE funded project identifying genes associated with chemoresponsiveness in ovarian cancer, as well as other future projects in this area.

The following table shows the current status of clinical data collection efforts, and includes data on participants enrolled in the DOD funded study as well as POCRC (ovarian SPORE) participants.

Table 1. Clinical Status Summary - October 2001			
Description	Total	% of Total	% of Category
Total women with Stage III/IV disease	121	72%	71%
Total women with Stage I/II disease	49	29%	29%
Total women with ovarian cancer to date	170	100%	100%
Total Stage III/IV greater > 9 months since surgery	110	65%	72%
Total stage I/II women > 9 months since surgery	43	25%	28%
Total women > 9 month since surgery	153	90%	100%
Total women stage III/IV characterized to date	88	52%	77%
Total women Stage I/II characterized to date	26	15%	23%
Total women all stages characterized to date	114	67%	100%
Total women in progress	39	23%	100%
Chemostatus Classification Detail			
Chemosensitive	41	24%	36%
Chemoresistant: Persistent	10	6%	9%
Chemoresistant: Progressive	12	7%	11%
<i>Subtotal: Specimens eligible for POCRC Project 1</i>	<i>63</i>	<i>37%</i>	<i>56%</i>
Chemosensitive: Ineligible for P1*	7	4%	6%
Chemoresistant: Persistent: Ineligible for P1**	1	1%	1%
Chemoresistant: Progressive: Ineligible for P1***	2	1%	2%

Chemoresistant: Recurrent	13	8%	11%
Insufficient Chemotherapy	13	8%	11%
No chemotherapy	15	9%	13%
<i>Subtotal: Specimens ineligible for POCRC Project 1</i>	<i>51</i>	<i>21%</i>	<i>45%</i>
Total	114	58%	100%
<i>*Ineligible due to neo-adjuvant chemorx</i>			
<i>**Ineligible due to Taxol resistance</i>			
<i>***Ineligible due to limited number of tx</i>			

Follow up process for data

With adherence to stringent enrollment procedures, very little direct participant follow-up for this study was needed, however on occasion follow-up was required if a patient had not fully completed enrollment forms or had not returned the study questionnaire. A protocol to address such circumstances was developed. In these situations, a written request, followed by one telephone call, is made by the Study Coordinator. The database system generated a report detailing a list of participants enrolled for at least 30 days, and for which there may be one or more pieces of enrollment information missing. A letter was mailed to the participant if their enrollment materials were incomplete or if the Core had not received their questionnaire within thirty days of enrollment. After fourteen days, a follow up call was made to the participant if she had not responded to the written request.

Collection and Preparation of Tissue and Blood Specimens

Specimen Collection

Recruitment and specimen collection for the ORCHID study was completed in Year 02. As of this report, 322 women consented to participate in the study, with successful ovarian tissue collections occurring on 257, of which 82 patients were diagnosed with ovarian cancer, 11 with tumors of Low Malignant Potential, 52 with benign disease, and 70 with no ovarian abnormalities. (32 collections were conducted with non-ovarian primaries or with no ovaries collected and 10 were blood only). The cancer cases include 12 patients with early stage disease, of which three are patients diagnosed with early-stage serous tumors.

A dedicated Tissue Collection Specialist was on hand to collect fresh and frozen tissue samples in addition to formalin-fixed and paraffin-embedded specimens. A Specimen, Collection, Processing and Storage protocol detailing collection techniques is included as Appendix Item M. Additionally, up to 50 cc of blood was collected and processed into sera,

plasma, and white blood cell and epithelial cell pellets. The variety of collection techniques allowed the Core to meet the needs of the projects within the program, as well as maximize use of the same tumors and develop the dedicated ovarian cancer repository.

QUEST Study Bloods

A portion of the blood specimens used as positive controls in the ORCHID study were obtained from women enrolled in the QUEST study. A total of 586 women were randomized to this study, of whom 292 were assigned to the ovarian cancer screening intervention arm. Women in this arm were consented for additional blood to be drawn for research purposes, including this study. To ensure the highest scientific integrity in analyses, the QUEST bloods were processed, inventoried and labeled in the same manner as bloods collected for the ORCHID study, and are stored in the same repository as the ORCHID bloods.

Specimen Allocation Procedures

Procedures for characterization and storage of ORCHID study specimens are being optimized in a manner that will suit future research project needs. Procedures to ensure that stored specimens are allocated for research in a manner which maximizes use of the resource have also been implemented. This includes a peer review of all specimens requested for use in outside studies. It also includes organization, aliquoting and distribution of samples for various biomarker analyses in a manner that preserved the integrity of the sample while utilizing limited amounts of specimen.

After Project needs were met, stored specimens could be made available to non-Project Investigators. Specimen accessibility is an important component of the repository. Epidemiologic and clinical data was collected and associated with stored tissue specimens, making the resource extremely valuable due to its large collection of specimens with detailed epidemiologic data and patient follow-up. Requests for access to specimens collected in the ORCHID study was previously handled by a combination of study Investigators, the FHCRC IRO and Human Specimens Committee. A separate specimen review committee implemented under the of the SPORE, reviews all requests for specimens. Membership includes ORCHID study investigators. Requests are reviewed for feasibility and scientific merit and may be granted in part or in full based upon the availability of specimens and priority assigned to the investigator's proposed work.

Specimen Transfer

For all specimen transfers, a report identifying specimens to be distributed is generated in the specimen inventory database. The investigator's name, laboratory location, and intended use is recorded in the database with the specimens (individually identified) to be sent to the research project. The Tissue Collection Specialist receives a copy of the pull and delivery report, removes the specimens from the repository, and packages the specimens securely for transport to the investigator's laboratory.

Upon receipt of the delivery, the investigator and Tissue Collection Specialist will review the contents of the delivery and check them against the printed report. Both sign a transmittal form confirming that the specimens listed were received in full and in satisfactory condition. The completion of this form and confirmation of delivery are stored in the specimen database.

and linked to the records of the specimens comprising the delivery. An example of this form is included with the specimen collection and tracking forms in Appendix N.

Histologic Characterization of Tissue and Blood Specimens

The Laboratory Core component of this study conducted a detailed review of all tissue specimens collected during surgery. These reviews allowed Investigators to rapidly identify appropriate cases for the projects and perform quality control of the tissue collection and processing. The procedures established in the ORCHID study are continued in the ovarian SPORE grant operations.

Dr. Nancy Kiviat was responsible for conducting pathology review of each formalin fixed tissue specimen collected during surgery for this study. The results of this characterization are coded and associated with each corresponding frozen tissue specimen in the specimen inventory database.

During the first two years of the study, histological examination was carried out on 209 collections with tissue and classified according to the World Health Organization (WHO) classification of ovarian tumors. Cases identified as tumors (benign or malignant) or other epithelial lesions were further characterized by immunohistochemistry.

We continue to analyze sera from study participants for CA125 levels. Inter-laboratory validation consisting of analyses conducted in two different laboratories using separate protocols were conducted. A collaborator in our SPORE grant, Dr. Irena King was responsible for analyses, and maintained good standing in the CAP (College of American Pathologists) validation program for the CA125II IRMA by Centocor. Additionally, Dr. King facilitated the measurement of MCSF by the in-house FHCRC cytokine laboratory personnel, reviewed completeness of the data prior to incorporation into the main database and discussed data interpretation with biostatisticians. The following table depicts the type and frequency of marker characterization for women whose specimens were collected for the ORCHID and QUEST studies.

Table 2. Summary of biomarker characterization and number of women with specimens in repository with an associated result.	
Assay Type	Number of women
Mcsf	413
Ca125	502
HE4 (6 dilutions)	278
Mesothelin (6 dilutions)	398
p53	488

her2/neu	197
tissue based HE4	67
tissue based Edg7	67
tissue based Mesothelin	67

Immunohistochemistry & Mutation Analyses

The Core laboratory conducted assays for the oncoproteins cerbB-2 and p53 from each malignant tissue and a fraction of all normal tissues. In addition, p53 DNA was isolated and analyzed for mutation in the p53 gene. The results generated by the Core laboratory were compiled and reported monthly to the study investigators. In addition to keeping investigators abreast of ongoing laboratory activity, this report served as a quality control measure that would reveal problems with screening assays or methods of tissue collection and processing.

Two hundred nine cases were characterized histologically according to the WHO classification of ovarian tumors. Of these, one hundred fifty tissues were characterized by immunohistochemistry. A panel of three antibodies was run on each case. Tissue reactivity was assessed using a monoclonal antibody directed against cytokeratin 8 (Becton Dickenson). To identify p53 overexpression, a monoclonal antibody which reacts with both the wild and mutant form of p53 was used (DAKO Corporation). Tumors that overexpressed the cerbB-2 oncogene product were identified using a polyclonal antibody (DAKO Corporation). The cases that were previously labeled as indeterminate were scored according to a scoring system developed by Dr. Allen Gown for breast carcinomas. This method provided good correlation between the cerbB-2 results obtained by immunohistochemistry and fluorescent in situ hybridization (FISH) detection of multiple gene copies. The staining intensity was graded on a 0 to 4+ scale. If normal internal or external controls were present, a subtracted score is obtained by subtracting the staining intensity of the normal control from that of the tumor. A tumor would be considered cerbB-2 positive if the subtracted score is greater than or equal to 2 or if the staining intensity of the majority of tumor cells are 3+ or greater. In all situations the staining pattern must be membranous and not cytoplasmic.

Forty cases consisting of normal, benign, and malignant tissues were characterized by a polyclonal antibody directed against the mutant form of EGFR (EGFRvIII). Problems were encountered with non-specific and high background staining with this antibody. This assay is on hold until a more specific antibody is developed.

The results of the histological and immunohistochemical characterization of the tumors, benign lesions and normal tissues are shown below:

Normal	# of Cases	P53 +	p53 -	cerbB-2 +	cerbB-2 -
Normal Appendix	1	0	1	0	1
Normal Cervix	6	0	2	0	2
Normal Colon	1	-	-	-	-
Normal Fallopian Tube	24	0	13	0	13
Normal Myometrium	9	0	1	0	1
Normal Ovarian Tissue	100	0	16	0	16
Normal Uterus	6	0	1	0	1
Corpus Luteum	2	-	-	-	-
Functional Cyst	15	-	-	-	-
Ovarian Fibroma	4	-	-	-	-

Benign Lesions	# of Cases	p53 +	p53 -	cerbB-2 +	cerbB-2 -
Benign Cyst, Not Paraovarian	8	0	8	0	8
Benign Cyst, Paraovarian	1	0	1	0	1
Endometriosis/ Endometriotic Cyst	8	0	4	0	4
Inflammatory Lesions	1	-	-	-	-

Neoplastic Other	# of Cases	p53 +	p53 -	cerbB-2 +	cerbB-2 -
Benign Brenner Tumor, Typical	1	0	1	0	1
Benign Dermoid Cyst	2	0	0	0	0
Thecoma	2	-	-	-	-

Serous Tumors, Benign	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Serous Adenofibroma	2	0	2	0	2
Serous Cysadenofibroma	3	0	3	0	3
Serous Cystadenoma	6	0	6	0	6

Serous Tumors, Low Malignant Potential	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Serous Carcinoma of LMP	6	0	6	0	6

Serous Tumors, Malignant	# of Cases	p53 +	p53 -	cerB-2 +	CerbB-2 -
Serous Carcinoma	42	27	15	7	35

Mucinous Tumors, Benign	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Mucinous Cystadenoma	5	0	5	0	5
Mucinous Cystadenofibroma	2	0	2	0	2

Mucinous Tumors, Low Malignant Potential	# of Cases	p53 +	p53 -	cerbB-2 +	CERBB-2 -
Mucinous Carcinoma of LMP	2	0	2	0	2

Mucinous Tumors, Malignant	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Mucinous Carcinoma	3	2	1	2	1

Endometrioid Tumors,	# of	p53 +	p53 -	cerbB-2 +	CerbB-2 -
----------------------	------	-------	-------	-----------	-----------

Malignant	Cases				
endometrioid Carcinoma	4	2	2	1	3

Clear Cell Carcinoma, Malignant	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Clear Cell Carcinoma	4	1	3	3	1

Neoplastic Other, Malignant	# of Cases	p53 +	p53 -	cerbB-2 +	CerbB-2 -
Adenocarcinoma, NOS	13	11	2	6	7
Unclassified Epithelial Tumor	5	3	2	1	4
Colonic Carcinoma	1	1	0	0	1
Cecal Adenocarcinoma	1	0	1	0	1
Endometrial Carcinoma	2	1	1	1	1

In the group of cases where primary and metastatic tissues were collected and characterized by immunohistochemistry, there were no differences in the p53 and cerbB-2 results between the primary and metastatic tumors (data not shown). None of the normal or benign lesions displayed overexpression of either p53 or cerbB-2.

Provision of Specimens to Projects

During the first year, Project 1 was provided with a total of 15 for the construction of cDNA libraries, with 31 specimens provided for first-generation membrane hybridization. This task has been completed and is described in greater detail in the Project 1 section of this report.

During Year two, Project 1 was supplied with a total of 54 specimens to hybridize with second-generation membranes. Originally, 105 specimens were to be provided to this project for second-generation hybridization. This number was revised to 75, and all specimens have been provided by the Core for this task.

During the first year, Project 2 was provided with provided with tissue specimens from two ORCHID study participants and 8 early stage specimens were obtained from the Gynecologic Oncology Group (GOG). In the second year, Project 2 was provided with 79 serum specimens of (benign and cancer of serous histology, and normal) from the study repository as well as obtained from the GOG. As serous tissue specimen inventory levels were not adequate at the time of allocation and additional 19 late stage snap frozen serous tissue specimens were obtained from the GOG for use in the SEREX analyses. During Year 3, an additional 30 blood specimens were provided to continue work on a second SEREX library. During Year 2, 179 specimens were provided for Project 2 to Dr. Mary L. Disis to assess antibody response.

Statistical Analyses and Design

Developing statistical algorithms for selecting over-expressed genes for subsequent efforts:

The underlying philosophy for analyzing the quantity of gene expression data on a relatively small sample was to provide a statistical filter that would reduce the number of candidate genes that were under investigation at one stage to a manageable level for the next more intensive level of investigation.

Our primary method ranks the gene candidates based on a modified t-statistic comparing the distribution of expression levels in ovarian cancer tissues to normal ovaries. This approach gives highest priority to genes with an average expression level in ovarian cancer tissues that are highly elevated over the average expression level in normal ovaries, as measured on a scale determined by the variability of expression levels. Initial analyses indicated that for a large proportion of genes of interest, the expression levels in normal ovaries are quite homogeneous. In ovarian cancer tissues, however, the expression values vary greatly. Using a pooled estimated of the variance in the usual t-statistic caused those genes with the greatest overdispersion in the cancers to receive a lower priority ranking, even when the discriminatory power should be very strong based on a visual examination of the distribution. To remedy this, we adopted a modified t-statistic where the estimate of the variance was based on the observed variability in the normal tissues alone. Though this statistic would not be considered "efficient" in the traditional sense because it does not use all of the information available, it provides a more relevant measure of difference in gene expression for our purposes. The

rankings from the RT-PCR studies using this approach is provided in Appendix O. A separate ranking was also performed based on the comparison to benign tissues but as the number of benign tissues analyzed to date is currently rather small, these values have not been formally incorporated into the selection process.

Statistical analysis of antibody response to Her-2/neu, p53 and c-myc

As described in Project 2, initial analyses of antibody responses to Her-2/neu and p53 levels were evaluated in two sets of normals, women with benign disease, borderline tumors, ovarian cancer and other cancers. C-myc performance was not evaluated. Both Her-2/neu and p53 are shown to be useful in discriminating ovarian cancers from normals in logistic regression analysis. After controlling for age and CA125 levels, however, Her-2/neu and p53 did not contribute significantly to the classification.

More definitive analyses were conducted during Year 3 to better assess the performance of antibodies to P53 and Her2/neu to distinguish between the sera of ovarian cancer cases and healthy women without ovarian cancer. We measured these markers in 68 cases of various histology, 40 specimens with benign ovarian disease, and 281 healthy women. The levels of the marker concentrations were compared between cases and controls, and their performance was measured using ROC curves. The performance of each marker alone was estimated, but also its performance to complement CA 125. The latter was done to fully evaluate the benefit of the marker, for a marker that performs less well on its own that CA 125 may still have benefit if it detects those cases missed by it. Summary performance of each marker alone and in conjunction with CA 125 is given below.

A ranking of each of these markers is given in the tables below. The p-values represents the result of a statistical test to determine if each marker, on its own, can detect cancer better than noise calculated from the Wilcoxon test. Each marker passes this test, but this test, although necessary, does not accurately reflect the needs of a marker for screening, which needs significant sensitivity at low ranges of specificity. One way to measure the markers ability is to measure the markers mean sensitivity over a specified range of specificity. The pAUC column below shows the average sensitivity of each marker when used over a range of specificity between 70% and 100%. This is often referred to as the normalized partial area under the ROC curve. A perfect marker has a score of 1.0. Each half of the table shows how well the marker performs when distinguishing cases from healthy controls (left half) and cases from benign controls (right half). The values of CA 125 is the best performing marker, with average sensitivity of 75%. Both Ab to P53 and Ab to H2n have approximately 30% sensitivity over that selected range of specificity

	VERSUS HEALTHY CONTROLS	
Marker	p-value	pAUC
CA 125	0	0.75
Ab to P53	0	0.33
Ab to H2n	0	0.36

	VERSUS BENIGN CONTROLS	
Marker	p-value	pAUC
CA 125	0	0.75
Ab to P53	0	0.30
Ab to H2n	0	0.29

Statistical analysis of Her-2/neu, p53 and c-myc expression in tissue:

This task is ongoing based on the data presented above (see Immunohistochemistry & Mutation Analyses), with evaluation continuing under the auspices of the SPORE grant.

Statistical analyses of select clones with clinical, epidemiologic and other laboratory data:

We have pooled much of the clinical epidemiologic and key laboratory data into an analytic data file. Appendix P presents a summary of the key factors by patient group: Normal, Benign, LMP (borderline) tumors, Ovarian Cancer, and Other Cancers. These groups differ in age, menopausal status and other factors that may be potentially related to biomarker levels. In our basic modeling to identify which marker or panel of markers best discriminate between ovarian cancers and normals, or ovarian cancers and benign disease, we will be cognizant of these differences and incorporate these factors into the models. As mentioned in Project 2, the analysis of Her-2/neu and p53 in conjunction with CA-125 levels suggested that these antibody responses provided only very modest improvement in the accuracy of a screen based on CA-125 and this did not reach statistical significance, despite the fact that the serum levels of these markers are not correlated.

We have identified several aspects that require further examination, and will be continued in our ovarian SPORE. In particular, we will obtain one or two other normal control groups from well-characterized cohorts having stored specimens available to us. This will give us an estimate of the distribution of these markers in women more representative of the general population that would be targeted for screening. Second, we will further analyze the incorporate factors such as tissue expression levels and other clinical features of disease to determine whether these antibody responses are related to specific subtypes of disease.

CONCLUSIONS

The development of a specimen repository with corresponding clinical and epidemiologic data required a substantial effort. This repository, from well characterized patients and with high quality, centralized pathology, and centrally determined laboratory measures provided a valuable resource for stimulating research on many aspects of ovarian cancer. Though initially targeted to early detection, this resource will be useful for testing hypotheses related to disease classification, prognoses, and response to therapy.

Our initial analyses of data from this project has revealed some of the challenges in the design and analyses for these types of biomarker development studies. We more clearly recognize the value of a well-characterized normal control groups that are representative of the population to be screened. We also note the need for further thinking on the appropriate normal controls for gene expression studies (e.g., normal contralateral ovaries, or normal ovaries removed for non-cancer indications), where truly tissue from individuals without any known pathology is very uncommon.

For analyses of gene expression data, statistical methods are being developed further. There are many groups around the world who are working on the different levels of this problem (e.g., sources of error, spot-finding, normalization). Dr. Schummer has collaborated with many of these to share his data and learn the results of their methods on these data. For purposes of early detection, as is our primary mission, we have assumed that a strong signal to noise ratio is needed in the gene expression level in tissue in order for this signal to be recognizable in a serum based assay. Under this assumption, we have used simple univariate approaches with raw data to identify and rank novel genes that can be investigated further. The success of this approach awaits the outcome of the next phase of this research.

In recognition of these challenges, the SPORE program developed a dedicated Informatics Core (IC) which builds upon the statistical, epidemiological and clinical foundation established in the DOD funded study. The Informatics Core is headed by Garnet Anderson, PhD, the primary statistician on this grant, and also includes Mary Anne Rossing, the DOD grant epidemiologist. This new core was developed to support the ongoing statistical analyses and data management efforts of investigators involved in the Pacific Ovarian Cancer Research Consortium (POCRC) projects and pilot studies, as a direct result of interpreting and analyzing ORCHID project data.

The advantages of an organized and comprehensive approach to informatics are several. The first advantage is ensuring consistent statistical analyses and reporting of laboratory-generated data. The second advantage is in providing scientific oversight of data associated with specimens in the specimen repository, to ensure complete and accurate reporting of such data, and provision of the data to investigators when required to meet scientific objectives. Another advantage is the ability to optimize the use of specimen resources by assuring adequate

statistical input in the design, and by identifying issues with overlap between separate investigations prior to the depletion of specific specimens. Additionally, this core will oversee the expansion of the POCRC information system, originally established under the DOD study, used to track specimens, to include management and correlation of laboratory generated data, while maintaining consistency with national common data elements initiatives.

The IC will be responsible for oversight, analyses, storage and reporting of data associated with specimens in the specimen repository. Most importantly, this core will also provide consultation and scientific assistance to laboratories developing potential ovarian biomarkers, and will be instrumental in assisting in the selection of markers for further development and validation. Consultation includes analyses of complex laboratory data generated by proteomics and high density array hybridization work, power calculations , and analysis of clinical trials. This service will ensure that collaborating ovarian cancer research projects are progressing as rapidly as possible.

KEY RESEARCH ACCOMPLISHMENTS

Project 1

- Construction of three unamplified and non-normalized cDNA libraries from normal ovaries, late stage ovarian carcinomas and metastatic ovarian carcinomas
- Generation of a cDNA membrane array consisting of 97,803 cDNA clones randomly selected from these libraries
- Interrogation of this array with probes from 30 tissues (normal and ovarian cancers) finding 17 genes (2 novel genes, 5 ESTs and 10 known genes) with marker potential
- Generation of a cDNA glass array consisting of 1390 genes selected from the membrane array with potential to code for marker genes
- Interrogation of this array with probes from 64 tissues (normal and ovarian cancer) finding 126 genes (8 novel genes, 30 ESTs and 88 known genes) with marker potential
- Expression validation of 78 genes by RealTime quantitative PCR, finding 15 marker genes
- ELISA test of SLPI on ovarian cancer patient sera reveals no elevated expression of SLPI protein in patient sera
- RT-PCR of MSLN and WFDC2 finds presence of transcript in epithelial cells from peritoneal washes of ovarian cancer patients but also of patients suffering from other malignancies and benign diseases.
- ELISA test of MSLN and WFDC2 on ovarian cancer patient sera suggests that both markers can add sensitivity and specificity to a CA125 assay.

Project Two

Task 1

- Fully operational and reproducible assay for detection of HER2 antibodies for use in final analysis.
- Fully operational assay for the detection of p53 antibodies.
- Construction of peptides which are dominant B cell epitopes of p53 and c-myc.
- Fully operational and reproducible assay for detection of HER2 antibodies using recombinant proteins.
- Development of conditions to detect peptide specific antibody responses by ELISA.
- Completed analysis of all ovarian cancer sera collected through the ORCHID study for HER2 and p53 antibodies (ug/ml) as well as analysis of control reference population (n=175) for HER2 and p53 antibodies.

Task 2

- Successful optimization of the signal-to-noise ratio for the SEREX protocol.
- Construction of a serous ovarian tumor cDNA library and a normal testes cDNA library.

- Development of an array based procedure that allows rapid evaluation of multiple phage clones with multiple serum samples.

Task 3

- Validation of the cDNA library and SEREX immuno-screening procedure by cloning the known ovarian tumor antigens p53 and NY-ESO-1.
- Successful use of the SEREX method to identify 26 candidate ovarian tumor antigens.
- Successful use of SEREX arrays to prioritize the 26 identified antigens on the basis of their immunogenicity across a panel of serum samples from 54 ovarian cancer patients and 20 normal controls.

Task 4

- Development of a reproducible ELISA protocol to assess serum antibody responses to NY-ESO-1, p53 in ovarian cancer patients.
- Production of Histidine-tagged recombinant versions of Ubiquilin-1, p53, KNSL6, HIS1 and HOXB for use in ELISA.
- Assembly of full-length cDNAs for other promising antigens, including FLJ20267, TOPO2a, RUVBL and DDX5.

Task 5

- A pooled dataset containing participant demographics, clinical characteristics, and selected laboratory values
- Analyses of the discriminatory power of antibody levels to p53 and H2N for distinguishing ovarian cancers from normal individuals both individually, jointly, and in combination with CA125 levels.

Core

- Development of a recruitment protocol and supporting documents
- Development of a tissue and serum repository containing biological specimens from 217 women
- Development of a study database that links epidemiological, clinical and laboratory data collected on all women enrolling in this project
- Provision of specimens to Projects 1 & 2
- Prioritization of gene expression from RT-PCR data for subsequent development into protein based serum assays

- Analyses of antibody responses to p53 and H2N as markers for classification purposes

REPORTABLE OUTCOMES

Project One

Publications:

1. Schummer M, Kiviat N, Bednarski D, Crumb GK, Ben-Dor A, Drescher C and Hood L (2000) Hybridisation of an array of 100,000 cDNAs with 32 tissues finds potential ovarian cancer marker genes, *Int. J. Biol. Markers*, 15 suppl. 1, 35
2. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M and Yakhini, Z (2000) Tissue classification with gene expression profiles. *Journal of Computational Biology* 7, 559-584
3. Keller A, Schummer M, Hood L, Ruzzo WL (2000) Bayesian Classification of DNA Array Expression Data. Technical Report, UW-CSE-2000-08-01, August, 2000.
4. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16, 906-14
5. Stone B, Schummer M, Paley PJ, Crawford M, Ford M, Urban N and Nelson BH (2001) MAGE-F1, a novel ubiquitously expressed member of the MAGE superfamily. *Gene* 267, 173-82
6. Schummer M, Kiviat N, Hood L, Drescher C, Ben-Dor A, Yakhini Z, McIntosh M, Siegel AF, Podolsky I, Hellström I, Hellström KE, Urban N (2001) Microarray-based gene profiling discovers potential ovarian cancer markers, submitted

Databases:

- orchidDB (FileMaker database holding the 100,000 membrane array clones and their hybridization signals across 57 conditions equaling 42 tissues)
- orchidGlassDB (FileMaker database holding the 1380 glass array genes and their hybridization signals across 64 tissues)

Funding:

- SPORE grant (for the identification of marker genes for chemoresistance in ovarian cancer).
- Pilot grants to the SPORE: 1) "Molecular genetic analysis of preneoplastic lesions of the ovary" with Andrew Godwin at Fox Chase Cancer Center, Philadelphia, PA; 2) "Development of ELISAs for the detection of potential ovarian cancer genes in sera" with Ingegerd and Karl-Erik Hellström at the Pacific Northwest Research Institute, Seattle, WA.

Project Two

Publications

1. Stone, B., Schummer, M., Paley, P.J., Crawford, M., Ford, M., Urban, N. and Nelson, B.H. 2001. MAGE-F1, a novel ubiquitously expressed member of the MAGE superfamily. *Gene* 267:173-182.
2. Stone, B., Crawford, M., Ford, M., Stewart, J., Thompson, L., Schummer, M., Paley, P.J., Urban, N. and Nelson, B.H. 2002. Serological Analysis of Ovarian Tumor Antigens Reveals a Bias Towards Antigens Encoded on 17q21. *In preparation*.

Abstracts

"Mapping the Immune Response to Ovarian Cancer for Screening and Therapy" Stone, B., Crawford, M., Ford, M., Stewart, J., Paley, P.J., and Nelson, B.H. 92nd Annual Meeting of the American Association for Cancer Research, March 24-28, 2001, New Orleans, LA.

Presentations:

- "Mapping the Immune Response to Ovarian Cancer for Screening and Therapy" Seminar, Pacific Ovarian Cancer Research Consortium, Fred Hutchinson Cancer Research Center, Seattle WA May 2000 (Brad Nelson, Ph.D.)
- Joint meeting, British Columbia Cancer Agency/University of British Columbia/University of Victoria, Victoria BC, June 2000. (Brad Nelson, Ph.D.)
- Seminar, Pacific Northwest Research Institute, Seattle WA, September 2000 (Brad Stone, Ph.D.)
- Annual Meeting, Society for Biological Therapy, Seattle WA, October 2000 (Brad Nelson, Ph.D.)

Patents:

U.S. Provisional Patent Application Nos. 60/278,253 and 60/278,237. Tumor Associated Antigens and Methods for Using the Same. Applicants: Brad Nelson (Seattle, WA) and Bradley Stone (Seattle, WA). Filing Date 3/24/2001. (see Appendix)

Active Funding:

NIH
1 RO1 CA82724-01
B.H. Nelson, P.I.

03/01/00-02/28/04
Direct total costs: \$1,442,014
Direct annual costs: \$339,580

"Novel Vaccine Targets for Early-Stage Breast Cancer"

The specific aims are:

1. To identify immunogenic proteins in early-stage breast cancer;
2. To select antigens for vaccine development on the basis of humoral and cellular immunogenicity in women with early-stage breast cancer.

NIH

8/1/00-7/31/02

1 R21 CA84359

Direct total costs: \$150,000

Brad Nelson, P.I.

Direct annual costs: \$75,000

"Immunologic Screening for Early-Stage Colorectal Cancer"

The specific aims are:

1. To classify 20 early-stage colorectal cancer patients as positive or negative with respect to serum antibody responses to a panel of known colorectal tumor antigens.
2. To determine whether patients who lack antibody responses to known tumor antigens instead respond to an undiscovered set of tumor antigens.

Completed Funding:

Morrison Trust

1/1/00-12/31/00

Brad Nelson, P.I.

Direct total costs: \$40,000

Direct annual costs: \$40,000

"A Novel Immunologic Blood Test for the Early Detection of Colorectal Cancer"

The specific aims are:

1. To identify a set of tumor proteins that commonly induce an antibody response in patients with early-stage colorectal cancer.
2. To determine the best combination of SEREX-defined tumor antigens to use for the detection of early-stage colorectal cancer.

Core

- Development of an ovarian specimen repository housing over 3000 individually identified specimens.
- Development of a participant database and specimen inventory tracking system.
- Funding of the 1999 ovarian cancer Specialized Program of Research Excellence by the NCI

Publications

1. McIntosh MW, Urban N. A Parametric Empirical Bayes Method for Cancer Screening Using Longitudinal Observations of a Tumor Marker. (*Biostatistics*).
2. McIntosh MW, Urban N, Karlan B. Generating Longitudinal Cancer Screening Algorithms for Novel Tumor Markers. (*Cancer Epidemiology Biomarkers and Prevention*)
3. Pepe MS, Longton G, Anderson G, Schummer M. Selecting Differentially Expressed Genes from Microarray Experiments (Submitted to Biometrics, 2001)

Invited Technical Talks/Poster sessions related to study:

- Combining CA 125, Mesothelin, HE4, Ab to Her2/neu, Ab to P53, and MCSF for detecting ovarian cancer, SPORE Workshop, Washington DC, July 2001.
- Methods for Selecting and Using tumor markers for ovarian cancer screening, SPORE Workshop, Washington DC, July 2001.

CONCLUSIONS

We have identified a large number of genes that are over-expressed in ovarian cancer tissue relative to the ovarian tissue obtained from women without cancer or ovarian pathology. In addition we have identified several oncogenic proteins that elicit antibodies detectable in the blood of some ovarian cancer patients. These discoveries are providing the foundation for ongoing work in early detection of ovarian cancer, funded by the NCI as part of a SPORE in ovarian cancer. Specifically, we are developing algorithms for using a panel of markers for ovarian cancer that tailors the use of the markers to the individual woman by accounting for change over time in each of the markers. We are continuing the process of evaluating the genes and gene products we have found for their likely contribution to the marker panel.

Our discoveries are expected to lead as well to work on the molecular characterization of ovarian cancer and a better understanding of ovarian cancer disease progression and biology. Several of the proteins we have found also have potential for therapeutic or prevention applications. Pilot studies to explore these possibilities are currently underway, and additional funding will be sought to continue our work toward improving the outcomes for women at risk for ovarian cancer. .

Final Report

Publications resulting from this Project (as of 10/2001):

1. Schummer M, Kiviat N, Bednarski D, Crumb GK, Ben-Dor A, Drescher C and Hood L (2000) Hybridisation of an array of 100,000 cDNAs with 32 tissues finds potential ovarian cancer marker genes, *Int. J. Biol. Markers*, 15 suppl. 1, 35
2. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M and Yakhini, Z (2000) Tissue classification with gene expression profiles. *Journal of Computational Biology* 7, 559-584
3. Keller A, Schummer M, Hood L, Ruzzo WL (2000) Bayesian Classification of DNA Array Expression Data. Technical Report, UW-CSE-2000-08-01, August, 2000.
4. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16, 906-14
5. Stone B, Schummer M, Paley PJ, Crawford M, Ford M, Urban N and Nelson BH (2001) MAGE-F1, a novel ubiquitously expressed member of the MAGE superfamily. *Gene* 267, 173-82
6. Schummer M, Kiviat N, Hood L, Drescher C, Ben-Dor A, Yakhini Z, McIntosh M, Siegel AF, Podolsky I, Hellström I, Hellström KE, Urban N (2001) Microarray-based gene profiling discovers potential ovarian cancer markers, submitted
7. Stone, B., Schummer, M., Paley, P.J., Crawford, M., Ford, M., Urban, N. and Nelson, B.H. 2001. MAGE-F1, a novel ubiquitously expressed member of the MAGE superfamily. *Gene* 267:173-182.
8. Stone, B., Crawford, M., Ford, M., Stewart, J., Thompson, L., Schummer, M., Paley, P.J., Urban, N. and Nelson, B.H. 2002. Serological Analysis of Ovarian Tumor Antigens Reveals a Bias Towards Antigens Encoded on 17q21. In preparation.
9. McIntosh MW, Urban N. A Parametric Empirical Bayes Method for Cancer Screening Using Longitudinal Observations of a Tumor Marker. (Biostatistics).
10. McIntosh MW, Urban N, Karlan B. Generating Longitudinal Cancer Screening Algorithms for Novel Tumor Markers. (Cancer Epidemiology Biomarkers and Prevention)
11. Pepe MS, Longton G, Anderson G, Schummer M. Selecting Differentially Expressed Genes from Microarray Experiments (Submitted to Biometrics, 2001)

Meeting abstracts (as of 10/2001):

"Mapping the Immune Response to Ovarian Cancer for Screening and Therapy" Stone, B., Crawford, M., Ford, M., Stewart, J., Paley, P.J., and Nelson, B.H. 92nd Annual Meeting of the American Association for Cancer Research, March 24-28, 2001, New Orleans, LA.

List of All Personnel Contributing to Project 1998-2001

Investigator/Staff	Role
Nicole Urban, ScD	Principal Investigator
Leroy Hood, PhD, MD	Molecular Biologist
Michel Schummer, PhD	Molecular Biologist
Nancy Kiviat, MD	Pathologist
Brad Nelson, PhD	Immunologist
Mary L. Disis	Immunologist
Brad Stone, PhD	Immunologist
Garnet Anderson, PhD	Statistician
Charles Drescher, MD	Gynecologic Oncologist
Mary Anne Rossing	Epidemiologist
Leona Holmberg, MD, PhD	Medical Oncologist
Beth Y. Karlan	Gynecologic Oncologist (Consultant)
Jane Kuypers, PhD	Laboratory Director
Janice Morihara	Immunocytochemistry Technician
Suepattra May	Project Coordinator
Steve Zeliadt	Project Coordinator
Sue Peacock	Statistical Research Associate
Karen Welsh	Systems Analyst Programmer (database)
Ted Cartwright	Systems Analyst Programmer
Paul Tittel	Data Coordinator
David Bednarski	Research Technician
Kaysey Orlowski	Tissue Collection Specialist
Susan Wilson	Intervention Specialist
Rachel Song	Desktop Publisher
Lisa Isozaki	Budget Management
Charles Trakarnsilpa	Network support
Marc Bruckner	Program Assistant
Nancy Myers	Program Assistant
Diane Koehnen	Administrative Support

Appendix A

Statement of Work

- 1. Project 1 Approved Statement of Work
- 2. Project 2 Approved Statement of Work
- 3. Core Statement of Work Status Table

III-B.6.d. Project 1. Statement of Work

Characterization of Genes Overexpressed in Malignant Ovarian Neoplasia by High Density Array Hybridization (Project Director: Lee Hood)

- Task 1.** Generation of representative cDNA arrays.
Months 1-4:
Six cDNA libraries will be generated from fetal, normal and benign, early stage and late stage neoplastic ovarian tissues.
These libraries will then be used to construct first generation solid phase membrane arrays containing 100,000 clones.
- Task 2.** Primary Characterization of Normal and Neoplastic Ovarian Tissue.
Months 5-10:
A. Hybridization of the first generation membranes with cDNA probes derived from 9 normal ovarian tissues (pre and post menopausal), 4 benign cystadenomas, 7 early stage and 7 late stage ovarian serous adenocarcinomas; 2 samples of bone marrow, and 2 samples of liver will serve as non-ovarian normal tissue controls.
B. Evaluation of hybridization results and selection of 2,000-3,000 genes overexpressed in malignant ovarian tissues.
C. Construction of second generation cDNA arrays containing these 2,000-3000 clones.
- Task 3.** Further Characterization of Gene Expression in Normal and Neoplastic Ovarian Tissue.
Months 11-15:
A. Hybridization of the second generation arrays with cDNA from tissues used in Task 2, plus 24 additional normal tissues (20 ovarian, 2 bone marrow, and 2 liver controls), 15 cystadenomas, 20 additional early and 20 late stage ovarian serous adenocarcinomas.
B. Evaluation of hybridization results and selection of ~400 genes that meet criteria for overexpression.
- Task 4.** Characterization of highly expressed genes associated with cancer.
Months 16-21:
A. Partial sequencing of the ~400 overexpressed ovarian cancer-associated genes identified in Task 3.
B. Homology searches through public databases to identify the ~200 *novel* genes for further analysis.
C. Multivariate statistical analysis of second generation screening data (as well as data from core studies and results of Specific Aim 1 of Project 2) for prioritization of candidate genes for further development.
D. Confirmation of tissue specific expression using RT-PCR and Northern blot techniques for ~20 leading candidate genes.
E. At least one high priority candidate cDNA will be provided to Dr. Nelson in Project 2 for evaluation of possible antibody responses in ovarian cancer patients and normal women.
- Task 5.** Final analyses and report writing.
Months 22-24:
Final analyses will be compiled and a final report and initial manuscripts will be prepared

III-B.6.d. Project 2 -- Statement of Work

Antibody Immunity to Cancer Related Proteins as a Serologic Marker for Ovarian Cancer (Project Director: Brad Nelson)

- Task 1:** Perform ELISA screens for p53, HER2/neu, Myc.
Months 1-24:
A. An ELISA based screen will be used to probe serum from ovarian cancer patients and control individuals for the presence of antibodies against the tumor associated proteins p53, H2N, and myc. It is anticipated to perform this set of tests on 350 cases per year.
- Task 2:** Determine SEREX baseline
Months 1-6:
A. Conduct ten serial absorptions on sera from three normal individuals and three ovarian cancer patients with known reactivities to one or more of the p53, H2N and Myc antigens.
B. Construct a cDNA expression library from pooled ovarian tumor samples.
C. Assess the quality of the library.
- Task 3:** Use SEREX to screen serum from ovarian cancer patients
Months 6-20:
A. Identify novel ovarian tumor antigens.
B. Prioritize the evaluation of novel ovarian tumor antigens. Statistical methods will be applied to identified antigens to determine if the discovery looks promising for translation into a test for use in the general public.
- Task 4:** Perform ELISA screens for Promising Candidates.
Months 18-24:
A. An ELISA based screen will be used to probe serum for the presence of antibodies against promising candidates that are identified by SEREX technology in Project 2.
B. Full length cDNA will be obtained and an ELISA based screen will be developed for at least one of the most promising overexpressed genes discovered through the HDAH in Project 1. **This task will not be completed by month 24.**
- Task 5:** Pool data for analysis.
Months 6-24:
A. All discovery data will be combined with data from the other labs through the coordination core. The Project Director will summarize on a routine basis the results and provide them to the Coordination Core for further interpretation and incorporation.

Statement of Work:

Tasks:

1. Define data collection instruments: Months 1-3
2. Develop data management and tracking systems: Months 1-9
3. Collect liver and bone marrow specimens: Months 1-4
4. Recruit surgery patients: Months 1-24
5. Collect tissue and blood specimens from surgery patients: Months 1-24
6. Collect blood specimens from QUEST participants: Months 1-24
7. Collect epidemiologic and clinical data: Months 1-24
8. Characterize histology of tissue specimens: Months 1-24
9. Perform tissue assays for p53, Her-2/neu, and c-myc: Months 3-24
10. Perform serum assays for CA-125: Months 3-24
11. Supply Project 1 with 6 specimens to construct cDNA libraries: Months 1-2.
12. Supply Project 1 with 30 specimens to hybridize with first generation membranes: Month 5.
13. Develop statistical algorithms for selecting over-expressed genes for subsequent efforts: Months 4-21.
14. Supply Project 1 with 105 specimens to hybridize with second generation membranes: Month 11.
15. Supply Project 2 with 10 ovarian tumor samples to construct cDNA expression library (Aim 2): Months 1-6:
16. Supply Project 2 with 600 blinded samples for antibody response analyses (Aim 1): Months 3-24.
17. Supply Project 2 with tissues samples from ovarian cancer cases and controls for SEREX analyses (Aim 2): Months 6-12.
18. Conduct statistical analysis of antibody response to Her-2/neu, p53 and c-myc (Project 2, Aim 1): Months 12-24
19. Conduct statistical analysis of Her-2/neu, p53 and c-myc expression in tissue: Months 12-24
20. Conduct statistical analyses of select clones with clinical, epidemiologic and other laboratory data: Months 21-24.

Major tasks and status listed in Core original Statement of Work

Function Associated with Task	Major Task	Progress
Patient Recruitment	1. Define Data Collection Instruments	Complete, Year 01
Data Warehousing Management	2. Develop data management and tracking systems	Complete, Year 01
Specimen Collection	3. Collect liver and bone marrow specimens.	Complete, Accumulated bone marrow only during Year 01
Patient Recruitment	4. Recruit surgery patients	Complete, Year 02
Specimen Collection	5. Collect tissue and blood specimens from surgery patients	Complete, Year 02
Specimen Collection	6. Collect blood specimens from QUEST participants	Complete, Year 03
Data management	7. Collect epidemiologic and clinical data	Complete, Year 03
Specimen Characterization	8. Characterize histology specimens	Complete, Year 03
Specimen Characterization	9. Perform tissue assays for p53, Her2/neu and c-myc	Complete, Year 03
Specimen Characterization	10. Perform serum assays for CA125	Complete, Year 03.
Specimen Distribution	11. Supply Project 1 with 6 specimens to construct cDNA libraries	Complete, Year 01.
Specimen Distribution	12. Supply Project 1 with 30 specimens to hybridize with 1 st generation membranes	Complete, Year 02.
Statistical Analyses	13. Develop statistical algorithms for selecting overexpressed genes.	Complete for DOD funding, Year 03. Ongoing under SPORE funding.
Specimen Distribution	14. Supply Project 1 with 105 specimens to hybridize with 2 nd generation membranes.	Initiated Year 02. In year 02 protocol modified and project needed 75 specimens. Complete, Year 02
Specimen Distribution	15. Supply Project 2 with 10 ovarian tumor samples to construct cDNA library	Complete, Year 01.
Specimen Distribution	16. Supply Project 2 with 600 blinded samples for antibody response analyses.	Complete, Year 03
Specimen Distribution	17. Supply Project 2 with tissue samples from ovarian cancer cases for SEREX analyses.	Complete, Year 02
Statistical Analyses	18. Conduct statistical analyses of antibody response to H2N, p53 and c-myc	Complete for DOD funding, Year 03. Ongoing under SPORE funding.
Statistical Analyses	19. Conduct statistical analyses of H2N, p53 and c-myc expression in tissue.	Complete for DOD funding, Year 03. Ongoing under SPORE funding.

Statistical Analyses	18. Conduct statistical analyses of select clones with clinical, epidemiological and other lab data.	Complete for DOD funding, Year 03. Ongoing under SPORE funding.
-------------------------	--	--

Appendix B

Project Timeline

- Project Timeline

10/30/01

Key:

x	Activity not yet implemented
x	Activity in process or complete
x	Activity in process and overdue

Appendix C
Project One: Figures

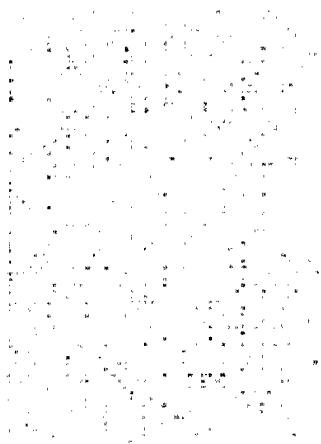


Figure 1 - Sample hybridization

Close view on 1/6 of a membrane containing 3456 colonies that was hybridized with a probe recognizing the vector portion of the cDNA. Where there is no signal, no colony grew. Overall, the number of colonies that did grow reaches 95%.

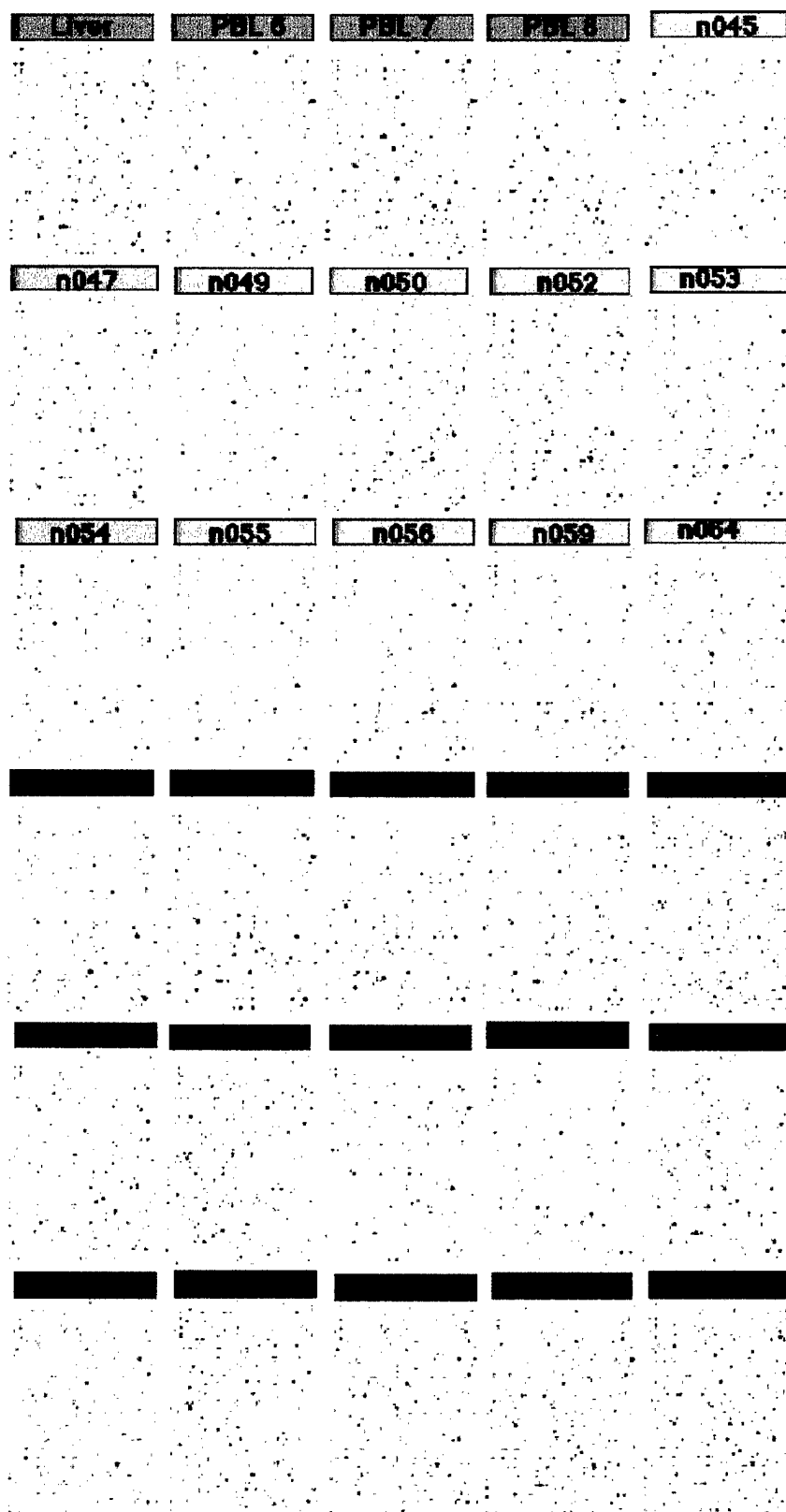


Figure 2 - Hybridization results

Displayed is one field containing 3456 colonies, replicated 30 times and hybridized with probes from 30 different tissues as indicated by the color. Although it may be possible to spot the most obvious differences and similarities in the hybridization pattern by eye, a computer-guided image processing is necessary to detect more subtle changes in expression.

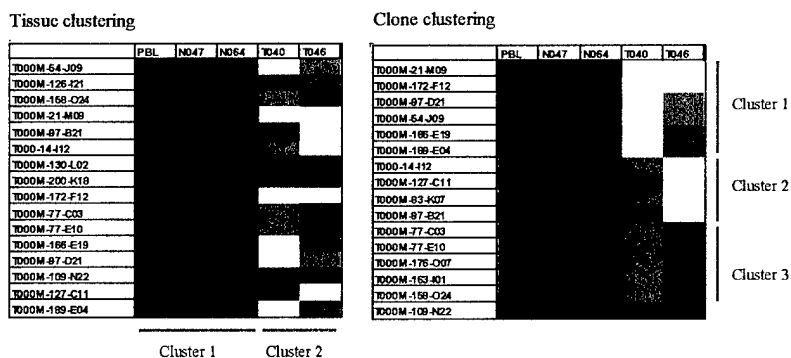


Figure 3 - Schematic explanation of the clustering

For better visual impression, the dataset is represented as a table and the values have been replaced by greyscale where white stands for high expression. Shown are 16 clones out of the 2651 (in the rows) and 5 hybridizations out of the 46 (in the columns): PBL (peripheral blood lymphocytes), two normal ovaries (N...) and two ovarian tumors (T...). In the left panel the tissues were clustered into two groups, one consisting of the normal ovaries and the PBL, the other consisting of the tumors. In order to select potential marker genes, the same clustering algorithm was repeated with a decreasing number of clones that would sort the tissues as nicely as displayed. The minimal number of clones that achieve this grouping are regarded as potential markers. In the right panel the clones were clustered into three groups. It is conceivable that members of a group are either clones representing the same gene or gene family or genes that share similar function or similar pathways. A clone that consistently clusters with a known tumor gene would be regarded as a potential marker gene. The small example shown here was applied to the full dataset as shown in Figure 4.

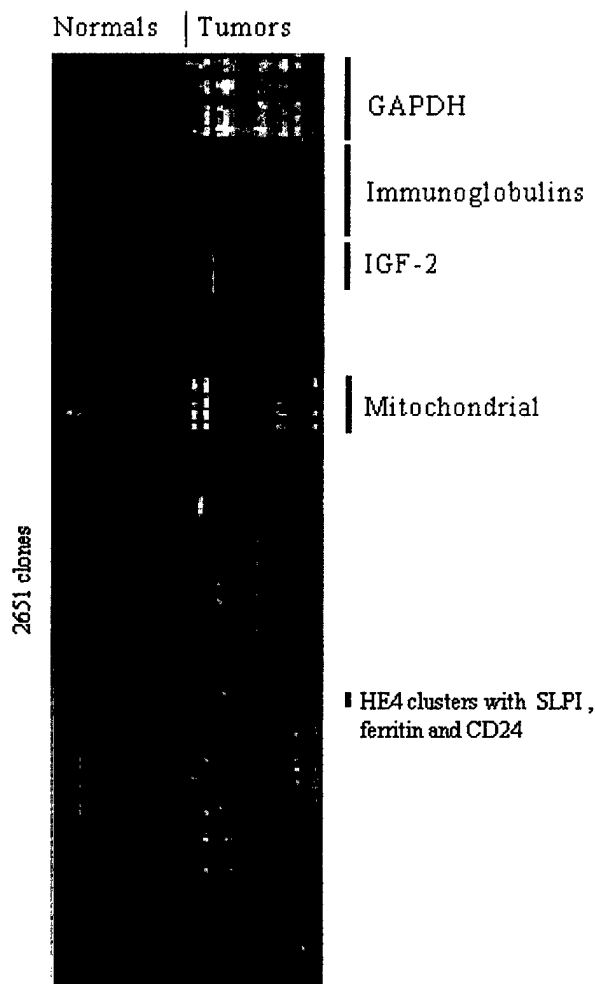


Figure 4 - Clone clustering on full dataset

Clone clustering performed on the full dataset of 2651 clones. The expression values are displayed as greyscale with white standing for high expression and black for a low one. The normal tissues (liver, PBL, normal ovaries) are shown on the left, the ovarian tumors on the right. Overall the expression of the normal tissues is lower than that of the tumors which reflects the selection criteria of these 2651 clones (low expression in normal tissues, high in tumors). In the present example the clones were clustered into 75 groups of varying size. The biggest groups consist to more than 80% of clones matching to GAPDH, immunoglobulins, IGF-2 and mitochondrial genes. Some of the smaller groups contain known tumor genes (such as CD24, ferritin and HE4) together with genes that were previously not known to be associated with tumors (such as SLPI and clones that do not match known sequences in the public databases). These clones were regarded as potential marker genes.

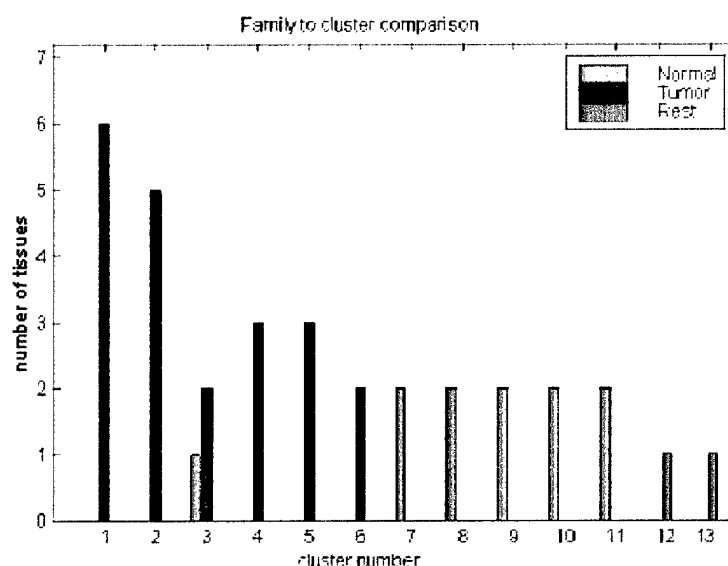


Figure 5 - Example of a tissue clustering result on the entire dataset

Displayed is a typical result for the leave-one-out tissue clustering analysis. The software generated 6 groups which - with the exception of one normal tissue - consist of tumors and five groups that contain only normal tissues. The duplicate and triplicate hybridizations of one tissue were treated as if they had been derived from separate tissues. As a result they either cluster in separate groups, which would be an indicator of low similarity, or they cluster in the same groups, indicating that they are indeed very similar to each other. Of the 7 tissues with repeated hybridizations, 5 have their replicates cluster in the same groups, one has two replicates in a "tumor" group and another replicates in a neighboring "tumor" group, and one has two replicates in a "normal" group and a single replicate in a "tumor" group. The groups 1-13 are formed from the following tissues: 1: hwb3, t037, t051, t051a, t040, t065; 2: t025, t060, t066, t044a, t044b; 3: n050a, t048, t044; 4: t046, t046a, t046b; 5: t063, t048a, t048b; 6: n039a, t043; 7: hpbl7, hpbl8; 8: n047a, n047b; 9: n050, n050b; 10: hliv2, hpbl6; 11: n056, n064; 12: t062; 13: t058. An "a" or a "b" behind the tissue name refers to the duplicate and triplicate hybridization.

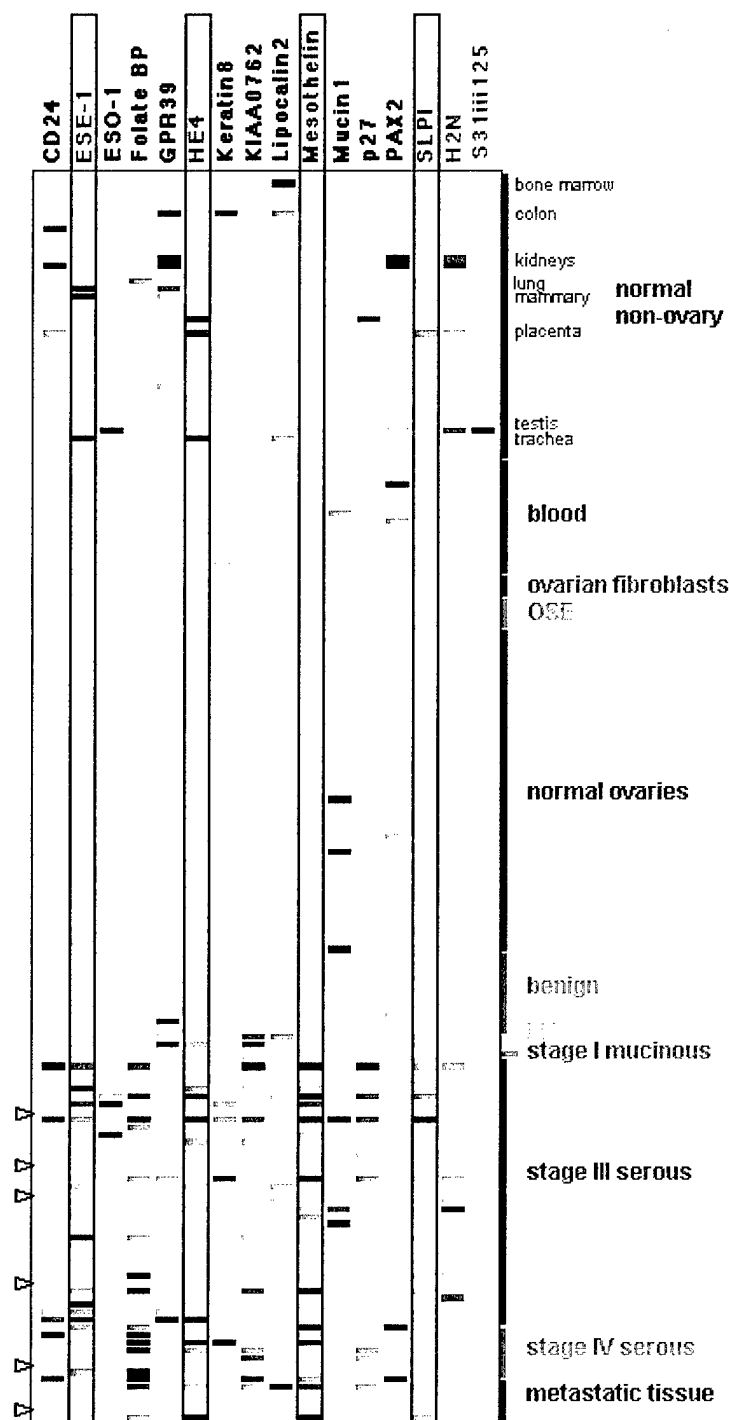


Figure 7 - RealTime data focusing on the expression of the marker genes in all tissues

The expression of 15 genes in 202 tissues was determined by RealTime quantitative PCR. Listed on the right are the tissues using the same colors employed throughout the report. The names of the genes are listed at the top. The expression values are expressed as greyscale bands with black standing for high expression and white for low. The four best performing genes are highlighted. The values are not normalized since normalization requires a gene or a group of genes with prior knowledge of their unchanged expression in the tissues tested. Since this is impossible, we have included in this panel the gene S3lilil125 which is expressed in all tissues shown, albeit with some variation. We would like to point out that had we normalized by the values of this gene, the overall expression pattern would still look the same with some bands being darker or lighter than otherwise. The open triangles on the left side mark tissues that show no elevated expression for either of the marker genes. H2N stands for Her2/neu.

CA125	p53	Her2/ neu	CD24	ESE-1	ESO-1	GPR39	HE4	Kera- tin8	Lipo- calin	Meso- thelin	Muc1n	p27	PAX2	SLP1	Tissue
U/ml	*	*	**	**	**	**	**	**	**	**	**	**	**	**	
21	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.8	0.0	0.0	n022
17	3	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.0	0.0	n029
7	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	n033
7	3	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n035
43	1	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	n041
9	1	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	n045
24	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	n047
	0	0	0.0	0.1	0.0	0.3	0.0	0.1	0.3	0.0	0.2	0.1	0.0	0.0	n049
20	5	3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n050
5	1	1	0.0	0.0	0.4	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n052
	0	0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n053
5	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n054
14	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	n055
7	0	0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n056
31	1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n057
	8	8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	n059
10	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	n064
	5	1	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.0	n082
24	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n083
9	5	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	n084
8	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n085
14	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	n087
8	0	0	0.3	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.1	0.2	0.2	n088
8	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	n089
12	0	0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	n090
14	0	0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.2	0.1	0.0	0.0	n092
11	0	0	0.0	0.2	0.0	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	n093
	5	3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.1	n094
17	0	0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	n095
17	0	0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.3	0.0	0.1	n096
	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.0	n097
64	0	0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.7	0.2	n100
	1	1	0.0	0.2	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.1	0.1	n102
10	0	0	0.0	0.1	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.3	0.6	0.0	n103
	5	3	0.0	0.2	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.1	0.0	0.0	n105
	0	0	0.5	0.5	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.1	0.0	0.1	n106
7	1	1	0.7	0.2	0.4	0.4	0.7	0.5	0.9	0.2	0.7	0.0	0.6	0.6	n108
61	0	0	0.2	0.1	0.0	0.6	0.2	0.3	0.0	0.1	0.2	0.1	0.0	0.3	n107
8			0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.0	n109
7	0	0	0.2	0.1	0.0	0.2	0.4	0.1	0.0	0.1	0.0	0.0	0.1	0.5	n109
22	0	0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	n109
4	0	0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.5	0.4	0.0	0.7	0.3	n109
	1	1	0.0	0.1	0.0	0.1	0.0	0.1	0.7	0.0	0.0	0.0	0.0	0.2	n115
81	5	5	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.2	0.4	0.0	n125
	0	0	0.4	0.2	0.0	0.9	0.7	0.2	0.1	0.4	0.8	0.0	0.5	0.3	n202
	1	1	0.0	0.3	0.3	0.0	0.1	0.7	0.4	0.8	0.3	0.3	0.5	0.3	n203
24	0	0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	n204
203	1	1	0.5	0.4	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.4	0.8	n205
174	1	3	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.2	0.0	0.3	0.2	0.7	n206
310	5	3	0.0	0.1	0.0	0.5	0.0	0.4	0.0	0.0	0.8	0.1	0.2	0.3	n057
	5	1	0.3	0.4	0.6	0.3	0.7	0.8	0.0	0.7	0.4	0.8	0.0	0.3	n019
170	1	1	0.8	0.2	0.8	0.3	0.6	0.1	0.1	0.4	0.8	0.6	0.5	0.7	n021
	5	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	n025
9	1	3	0.2	0.1	0.0	0.8	0.6	0.1	0.5	0.5	0.1	0.3	0.1	1.0	n031
	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	n039m
	8	8	0.5	0.2	0.0	0.3	0.4	0.5	0.0	0.2	0.6	0.4	0.0	0.6	n043
342			0.4	0.2	0.0	0.0	0.0	0.4	0.3	0.0	0.2	0.0	0.0	0.5	n049
			0.0	0.0	0.0	0.3	0.1	0.1	0.3	0.0	0.0	0.3	0.0	0.2	n051
	5	3	0.3	0.2	0.0	0.3	0.7	0.2	0.3	0.3	0.4	0.4	0.3	0.4	n060
14	1	3	0.1	0.2	0.0	0.0	0.7	0.9	0.0	0.9	0.2	0.5	0.0	0.4	n061
18	1	1	0.1	0.0	0.0	0.9	0.2	0.3	0.1	0.4	0.1	0.3	0.6	0.7	n065
18	0	0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.5	0.0	0.5	n066
91	0	0	0.5	0.4	0.0	0.8	0.1	0.4	0.5	0.6	0.0	0.4	0.8	0.7	n086
	0	0	1.0	1.1	0.0	0.3	0.4	0.6	0.0	0.8	0.0	0.7	0.0	1.4	n098
36	0	0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	1.1	n101
51	5	3	0.1	0.2	0.0	0.7	0.3	0.2	0.2	0.0	0.1	0.1	0.7	1.3	n104
	0.9	0.2	0.0	0.0	0.0	0.8	0.6	0.5	0.3	0.2	0.5	1.6	4.8	2.0	n107
	5	3	0.1	0.2	0.0	0.4	0.2	0.4	0.3	0.2	0.4	0.2	1.9	2.4	n108
344	0	0	0.1	0.4	0.0	0.1	0.8	0.2	0.0	0.5	0.4	0.0	0.0	0.2	n109
	5	5	0.0	0.6	0.0	0.6	0.2	0.4	0.0	0.7	0.4	0.6	0.0	1.1	n110
	5	3	0.6	0.5	0.0	0.6	0.7	0.8	0.3	0.0	0.3	0.5	0.3	0.3	n111
	0	0	0.9	0.6	0.0	0.5	0.6	0.4	0.0	0.0	0.0	0.0	0.7	0.7	n112
	5	1	0.3	0.6	0.0	0.1	0.6	0.6	0.9	0.0	0.3	0.4	0.0	0.3	n113
306	8	8	0.2	0.1	0.0	0.7	0.5	0.7	0.0	0.0	0.0	0.1	0.0	0.5	n114
	1	5	0.2	0.1	0.0	0.8	0.7	0.6	0.5	0.0	0.6	0.0	0.0	1.8	n116
51	5	3	0.0	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.1	n117
	5	3	0.8	0.8	0.0	0.4	0.3	0.2	0.2	0.0	0.8	0.0	0.2	2.1	n118
	1	1	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.6	0.1	0.0	0.1	n120
	1	3	0.4	0.5	0.0	0.4	0.7	0.8	0.2	0.2	0.2	1.0	0.0	1.5	n122
	5	3	0.5	0.5	0.0	0.0	0.5	0.8	0.0	0.1	0.0	0.1	0.0	1.0	n124
8	1	1	0.9	0.9	0.0	0.7	0.3	0.3	0.3	0.0	0.7	0.6	0.4	2.3	n206
	0	0	0.6	0.9	0.0	0.5	0.6	0.0	0.4	0.0	0.5	0.8	0.0	1.1	n119
	1	1	0.5	0.4	0.0	0.2	0.3	0.3	0.6	0.0	0.5	0.8	0.0	0.6	n040
63	1	1	0.1	0.4	0.1	0.2	0.0	0.2	0.0	0.0	0.4	0.1	0.0	0.1	n046
	0	0	0.0	0.0	0.0	0.2	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.0	n055
	0	0	0.0	0.4	0.0	0.6	0.5	0.5	0.5	0.2	0.2	0.0	0.0	1.6	n123

Figure 8 - Combined protein/transcript data focusing on the tissues of which there is CA125 information available

CA-125 serum levels of the ovarian cancer patients and the controls paired with the tissue protein levels of p53 and Her2/neu, listed side-by-side with the transcript levels of selected potential marker genes found in this study.

The patient diagnosis / tissue type is listed in the rightmost column (colors are the same as used in Figure 7). Values are overlaid with color for easier identification: CA-125: 0-29 U/ml (turquoise), 30-99 U/ml (faint red), 100-399 U/ml (red), over 400 U/ml (dark red), white: not done. * p53 and Her2/neu: 0, assay not run; 1, no overexpression (turquoise); 3, uninterpretable (light red); 5, intermediate overexpression (red); 6, high overexpression (dark red); 8, assay will not be run. The RealTime quantitative PCR values were normalized by the average expression of each gene in all tissue in order to have the values in each column on the same scale. ** 0-0.1, no expression (white), 0.2-0.9 weak expression (light red), >1.0, high expression (red). Ovarian cancer patients with CA-125 levels below 30 U/ml that have high levels of one or more of the newly found markers are labeled with a black dot after the tissue name.

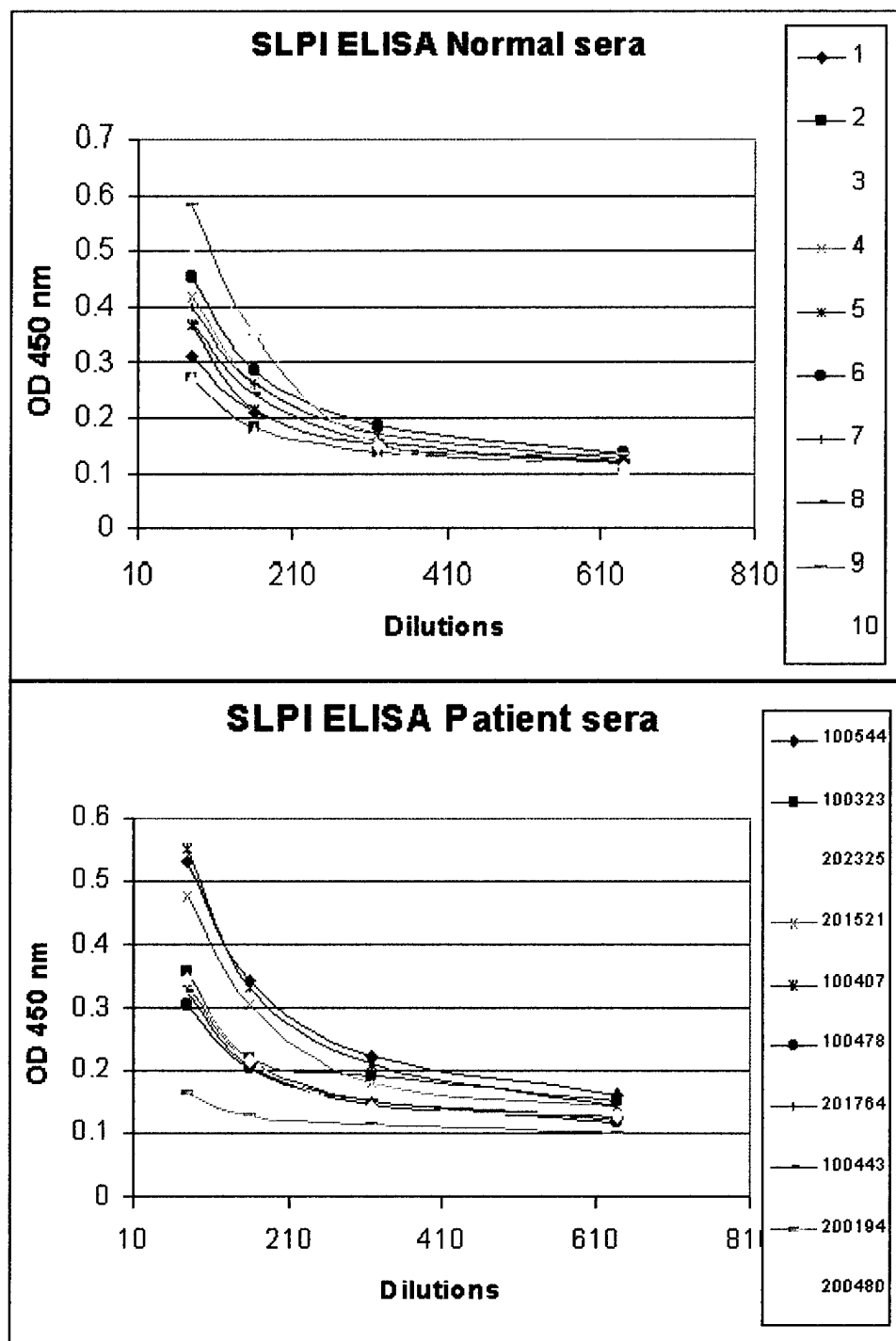


Figure 9 - SLPI ELISA on sera from 10 ovarian cancer patients and 10 controls

Top: sera from 10 normal controls, bottom: sera from 10 ovarian cancer patients whose tissues showed overexpression of SLPI message as assayed by RealTime quantitative PCR (see Figure 7). There is no difference in protein levels between the two groups.

Appendix D
Project 1: Related Publications

- “Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas
- “Tissue Classification with Gene Expression Profiles”
- Abstract: “Hybridisation of an array of 100,000 cDNAs with 32 tissues find potential ovarian cancer marker genes”
- “Microarray-based gene profiling discovers potential ovarian cancer markers”

Comparative hybridization of an array of 21 500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas

Michèl Schummer ^{a,*}, WaiLap V. Ng ^a, Roger E. Bumgarner ^a, Peter S. Nelson ^a,
Bernhard Schummer ^b, David W. Bednarski ^a, Laurie Hassell ^a, Rae Lynn Baldwin ^c,
Beth Y. Karlan ^c, Leroy Hood ^a

^a Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA 98195, USA

^b Institut für Pharmakologie und Toxikologie, Fakultät für Klinische Medizin der Universität Heidelberg,
Maybachstr. 14-16, 68169 Mannheim, Germany

^c Department of Obstetrics and Gynecology, Cedars-Sinai Medical Center, University of California, Los Angeles, School of Medicine,
Los Angeles, CA 90048, USA

Received 28 January 1999; received in revised form 2 July 1999; accepted 28 July 1999; Received by I. Verma

Abstract

Comparative hybridization of cDNA arrays is a powerful tool for the measurement of differences in gene expression between two or more tissues. We optimized this technique and employed it to discover genes with potential for the diagnosis of ovarian cancer. This cancer is rarely identified in time for a good prognosis after diagnosis. An array of 21 500 unknown ovarian cDNAs was hybridized with labeled first-strand cDNA from 10 ovarian tumors and six normal tissues. One hundred and thirty-four clones are overexpressed in at least five of the 10 tumors. These cDNAs were sequenced and compared to public sequence databases. One of these, the gene *HE4*, was found to be expressed primarily in some ovarian cancers, and is thus a potential marker of ovarian carcinoma. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Cancer maker; DNA array; Differential expression; HE4

1. Introduction

Ovarian cancer is the leading cause of gynecological cancer death in the United States. The American Cancer Society estimates that in 1998, some 25 400 women will develop ovarian cancer and 14 500 will die from it (American Cancer Society, 1998). The overall 5 year survival rate is about 46%, and has remained essentially unchanged for 25 years. Ovarian cancer is ranked fifth in cancer mortality among women, and raises concerns both with women and physicians because of its generally poor prognosis. Cancers diagnosed at an early stage have a 5 year survival rate of 92% in contrast to a 25%

5 year survival rate for patients with disseminated disease at diagnosis. Seventy-five per cent of epithelial ovarian cancers are diagnosed at advanced stages. This is in part due to the lack of symptoms early in the disease course, and the absence of a sensitive and specific screening test for early disease detection. Currently available ovarian cancer markers such as CA-125 are neither sensitive nor specific enough for population screening to detect early, treatable ovarian cancers (Jacobs et al., 1993).

We describe the use of 'high-density cDNA array hybridization' (HDAH) to identify transcripts that show high expression levels in ovarian cancer tissues as compared to ovarian surface epithelium (OSE). This technology has been used in a variety of experiments to identify transcripts (Schena et al., 1998), whose expression patterns differ in two tissues (e.g. normal and cancer). Our objective is to find (1) transcripts that are overexpressed in tumor as contrasted with normal ovarian tissue and (2) cDNAs encoding proteins that could be useful diagnostic markers (e.g. secreted or cell-surface pro-

Abbreviations: bp, base pair(s); cDNA, copy DNA; EST, expressed sequence tag; HDAH, high-density array hybridization; HE4, human epididymis gene 4; kb, kilobase(s); nt, nucleotide(s); OSE, ovarian surface epithelium; PBL, peripheral blood lymphocytes; RT-PCR, reverse transcription polymerase chain reaction.

* Corresponding author. Tel.: +1-206-616-5117;
fax: +1-206-685-7301.

E-mail address: kikjou@u.washington.edu (M. Schummer)

teins). Two general types of assays are possible: (1) protein assays for secreted proteins or on the surface of cells that metastasize into the circulation, and (2) PCR assays from genes uniquely expressed in blood-borne (or ascites-borne) tumor cells. Hybridizing 21 500 randomly selected cDNAs from normal and neoplastic ovarian tissues with probes from 10 ovarian tumor and six normal tissues, we identified 134 clones with higher expression signals in ovarian tumors as opposed to normal tissues. These clones were sequenced, and in some cases, their expression pattern was confirmed by RT-PCR and Northern blot analysis. The expression pattern of one of these clones, *HE4*, suggests that it may be a potential candidate diagnostic marker for ovarian cancer.

2. Materials and methods

2.1. Tissues and cells

We used the following tissues for our experiments: ovarian surface epithelium short-term culture (Karlán et al., 1995), early passages (OSE); normal ovary consisting of primarily stromal cells (N002, N005, N006, N019 and N035); two benign ovarian tumors (T017B, an endometrioid polyp, and T018B, a serous cystadenoma); one borderline early stage serous carcinoma, LMP (T028L); late-stage, high-grade papillary serous ovarian adenocarcinomas (T001–T006, T008–T011, T014–T016 and T021); two early-stage ovarian adenocarcinomas (one serous: T007 and one mucinous: T037); one late-stage, high-grade serous ovarian adenocarcinoma post-chemotherapy (T012); two late-stage, high-grade serous ovarian adenocarcinoma with massive metastases (T013M and T026M); peripheral blood lymphocytes (PBL1 and PBL2); Fetal ovaries: pool of 25 fetal ovaries (52–103 days); bone marrow, cerebellum, kidney, liver and placenta (Clontech, Palo Alto, CA). In order to minimize the effect of variance in tissue collection on the RNA quality and hence the hybridization patterns, we ensured that tissue collection would adhere to the following guidelines. After surgery, a tissue section was taken for the pathologist's examination and an adjacent section was snap-frozen in liquid nitrogen. All ovarian tumor tissue specimens were examined for their tumor cell content (which was above 80%) and the absence of necrosis. RNA preparations of all tissues or cell cultures were performed using the Trizol method (Life Technologies, Grand Island, NY). Poly(A)⁺ RNA was prepared using a mRNA purification kit (Stratagene, La Jolla, CA). Tissue samples of 200–400 mg of tumor were used for RNA preparation. We have found that samples of less than 200 mg do not yield sufficient RNA for our analysis. The integrity of total RNA was determined by visual inspection of the

28S and 18S ribosomal bands to ensure that degraded samples that might give a different expression profile than intact RNA were not used.

2.2. Minipreparation of 21 500 ovarian clones

Five cDNA libraries were created from ovarian tissues and cell cultures (OSE, T007, T008, T010 and T012) using the ZAP-cDNA synthesis kit (Stratagene). Examining the cDNA clones using PCR, the insert sizes were found to average between 1.2 and 1.5 kb. From each library, 96 clones were randomly chosen, sequenced and analyzed by similarity analysis against the non-redundant and EST database. The low number of mitochondrial and ribosomal sequences, the limited number of clones with no insert, and the significant cDNA diversity indicated that the libraries were of high quality. Using a 96-deep-well plate-based miniprep assay (Ng et al., 1996), we picked 21 500 transformants (8600 from the OSE cDNA library and 3225 each from the four tumor cDNA libraries), extracted the cDNAs and transferred them to 384-well microtiter plates.

2.3. Dotting the 21 500 clones onto nylon membranes

Using a hand-held arraying tool with a 384-pin printhead developed in our laboratory (Schummer et al., 1997), we dotted the 21 500 cDNAs onto 16 sets of 14 nylon membranes of 7.5 × 12 cm, which held each of the 1536 clones. The cDNA was denatured and immobilized on the membrane as previously described (Schummer et al., 1997).

2.4. Labeling and hybridization protocol

Each set of membranes was hybridized with a complex probe consisting of ³²P-labeled first-strand cDNA. Briefly, 5 µg of poly(A⁺) RNA or 30 µg of total RNA were reverse-transcribed using Superscript II reverse transcriptase (Life Technologies) and oligo-dT₁₂ primers with 30 µCi of alpha-³²P-dCTP (3000 Ci/mmol) and unlabeled dATP, dGTP, dTTP at 1 mM each; after 20 min, unlabeled dCTP was added to a final concentration of 1 mM, and the reaction was continued for another 40 min. This unpurified probe was hybridized to 12 membranes under conditions described previously (Schummer et al., 1997). The membranes were washed at increasing stringency (20 min, 2 × SSC, 0.5% SDS, RT; 20 min 0.5 × SSC, 0.5% SDS, 65°C; 2 × 20 min, 0.2 × SSC, 0.5% SDS, 65°C).

2.5. Software for spot detection

After hybridization and washing, the membranes were exposed to a phosphor storage screen, and the hybridization patterns were captured as 16-bit TIFF

images using a PhosphorImager (Molecular Dynamics, Sunnyvale, CA). Nine nylon membranes were imaged simultaneously on a 35 × 45 cm screen. The resulting file was processed using a software package developed in our laboratory. The TIFF image was split into nine smaller images, each representing one of the arrayed membranes. Briefly, the user defined the outer dimensions of each membrane by placing a cursor into each of the upper left, upper right and lower right corner of each of the nine array images. Subsequently, the computer superimposed a grid, approximating the positions of the 1536 dots. By five passes of center-of-mass finding, the computer determined the exact center of each of the 1536 dots. It integrated the area of an experimentally determined number of pixels around each center that covered the area of the largest hybridization signal present on the membranes. The intensities of all pixels in the area were integrated. Local background was calculated by choosing one pixel with the lowest intensity out of four pixels situated halfway between one dot and its four diagonal neighbors. Both values were stored in a tab-delimited text file together with the coordinates of the spot on the array.

2.6. Single pass 5' sequencing, database analysis and sequence comparison

Sequencing was performed on plasmid DNA and PCR products using previously described methods (Ng et al., 1996). The single-pass sequences were edited to remove vector and poly(A) sequences. Edited sequences were compared with those in the EST (dbEST) and non-redundant nucleotide and protein databases (GenBank) at the National Center for Biotechnology Information (NCBI) using the Baylor College of Medicine Search Launcher batch client server 'Search Launcher' (<http://www.hgsc.bcm.tmc.edu/SearchLauncher/>). Nucleotide sequence comparisons were carried out using BLASTN. Comparisons of conceptual protein translations were performed using the program BLASTX with BEAUTY sequence annotation enhancement. Each clone was categorized as to known gene homology, EST homology, or novel.

2.7. RT-PCR

Clones determined by to be differentially expressed by array analysis were confirmed by single tube RT-PCR, which has been shown to be a highly sensitive measure of transcript abundance (Schummer et al., 1998). Two primers, with a base pair length of 20–24 and with T_m s between 64 and 66°C, were designed for each gene. The distance between the primers was 420–660 bp. RT-PCR (Titan[®], Boehringer Mannheim, Mannheim, Germany) was performed with 200 ng of total RNA according to the manufacturer, with the

following cycles: 30 min at 50°; 2 min at 94°; 10 cycles of 30 s at 94°, 30 s at 60°, 45 s at 68°; 12–25 cycles of 30 s at 94°, 30 s at 60°, 45 s at 68° (with elongation of 5 s for each cycle); 7 min at 68°. For each gene, the logarithmic phase of amplification was determined prior to the Titan[®]-PCR. The individual reactions were run on a 1% agarose gel stained with SYBR-Green at 500 × diluted concentration for 1 h and scanned on a FluorImager (Molecular Dynamics, Sunnyvale, CA). For each gene and tissue, four identical reactions were performed.

2.8. Northern blot

A HE4 PCR product of 500 bp was cloned into a pCR2.1 vector using the TA cloning kit (Invitrogen, San Diego, CA). A digoxigenin-labeled riboprobe was prepared from this vector using a Genius RNA DIG labeling kit (Boehringer Mannheim, Germany). The probe was hybridized overnight at 68°C in DIG Easy Hyb buffer and washed in 2 × SSC, 0.1% SDS for 15 min at room temperature; 2 × SSC, 0.1% SDS for 20 min at 68°C; and 0.1 × SSC, 0.1% SDS for 2 × 15 min at 68°C. The hybridized RNA was visualized using the DIG detection kit (Boehringer Mannheim), and the membrane was exposed to X-ray film for 15 min.

3. Results and discussion

3.1. Evaluation of high-density filter hybridization

Tissues comprise many different cell populations. Each type of cell in a tissue exhibits its particular gene expression pattern. Since most ovarian tumors arise from epithelial cells, the comparison of tumors against ovarian surface epithelium should provide a useful comparison. Two qualifications must be made: (1) ovarian surface epithelial cells in a short-term culture will probably have some differences in expression patterns from in-vivo ovarian epithelial cells, and (2) tumors may have intermixed normal cells from the ovary. In order to detect genes that are overexpressed in one cell type or tissue versus another, one needs to know the limitations of the detection system, notably (1) the upper and lower limits of detection (signal-to-noise ratio) which — translated into the number of mRNA molecules detectable per cell — should be suitable for the proposed study, and (2) the measured level of variation in signal intensity on identical membranes interrogated with identical probes. The latter will determine a factor above which overexpression can be regarded as significant.

3.1.1. Determination of detection limit and dynamic range

The sensitivity of the array technology determines the number of detectable mRNA molecules in a cell. In order to determine the mean signal-to-noise ratio, we hybridized 14 identical arrays containing 1536 identical cDNAs coding for the green fluorescent protein (*GFP*) with first-strand cDNA probes made from human liver poly(A)⁺ RNA in which a *GFP* mRNA was added in decreasing concentrations (14 different concentrations ranging from one transcript in 200 to one in 20000). As depicted in Fig. 1, the probe with the highest *GFP* concentration yielded a mean value of 8300 ± 416 dpm (decays per minute) per pixel, and the mean background value was determined as 90 ± 18 dpm/pixel. With background subtraction, this represents a dynamic range of 456 (background-subtracted signal divided by background fluctuation: $8210/18=456$) or 2.5 orders of magnitude. We established a lower limit of sensitivity of 1 *GFP* RNA in 20000 liver RNAs, a result similar to those in other studies (Piétu et al., 1996). Based on an estimated 10^5 – 10^6 transcripts per average eukaryotic cell (Bishop et al., 1974), the membrane-based HDAH can detect a minimum of between five and 50 mRNA molecules in a cell and a maximum of 500–5000. The lower limit falls in the low to medium class of transcripts, and the upper limit lies in the highly expressed gene class (Zhang et al., 1997). This detection range should be sufficient for the identification of overexpressed genes.

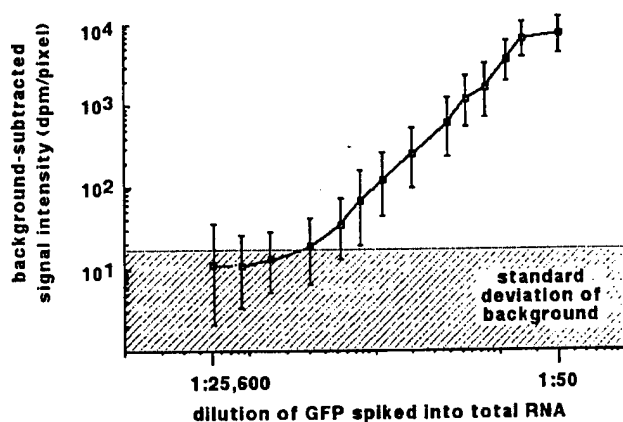


Fig. 1. Determination of the linearity of the hybridization signal. Fourteen replica membranes with 1536 *GFP* cDNAs each were hybridized with first-strand cDNA made from poly(A)⁺ RNA from human liver with *GFP* mRNA spiked into it in a twofold serial dilution starting with a 1:50 dilution. The signal intensity is measured as the average of all pixels of all 1536 signals. Displayed are the background-subtracted intensities with their standard deviations. The resulting curve is linear over 2.5 orders of magnitude. The standard deviations increase with decreasing signal-to-noise level. The shaded area indicates the standard deviation of the background. The background intensity averaged at 90 ± 18 dpm/pixel, and the highest intensity averaged at 8300 ± 416 dpm/pixel.

3.1.2. Normalization of the hybridization signals

In order to compare hybridization signatures of two identical membranes that have been hybridized with different probes in two separate incubations, one needs to normalize the signals to a standard. Although we adhere to a strict protocol, slight variations can be introduced by minute differences in probe labeling, probe purification, hybridization and wash conditions and exposure time. We normalized the background-subtracted intensities of one membrane by setting the median to 1. Assuming that among the 1536 clones present on one membrane, the majority does not alter its expression (Zhang et al., 1997), we believe that this is justified.

3.1.3. Determination of variation in signal intensity

Two factors influence the accuracy of the hybridization detection for one particular cDNA on a membrane: the amount of cDNA on the membrane (governed by the dotting procedure) and the amount of labeled cDNA that remains bound to the target cDNA on the membrane after hybridization (governed by the efficiency of the probe labeling reaction and the hybridization and washing kinetics). We determined the variation of amounts of DNA spotted by our arraying tool to be $\pm 14\%$ (data not shown). Since the probe consists of a complex mixture of cDNAs, the arrayed DNA is in vast excess of the probe cDNA, and thus the variations caused by the spotted cDNA can be regarded as negligible. In order to assess the probe-to-probe variance, we hybridized four replica membranes containing 1536 ovarian cDNAs with four ³²P-labeled first-strand cDNA probes independently generated from one batch of total RNA prepared from liver tissue. We compared the background-subtracted intensities of one cDNA across the four membranes and calculated the standard deviation, thus generating 1536 values. We ranked the clones by their expression and determined three means of standard deviations, one for the upper, the middle and the lower third, corresponding roughly to the high, medium high and low, expression categories of transcripts. The mean of the standard deviations amounted to $\pm 15\%$, $\pm 24\%$ and $\pm 40\%$ respectively, which averages to $\pm 26\%$ for all clones. Using the following equation, we calculated the threshold value for a ratio to be regarded as significant: $[1 + \text{standard deviation}] / [1 - \text{standard deviation}]$. In order to be above this threshold of significance, a highly expressed gene needs to display a ratio of 1.35, a medium expressed gene a value of 1.63 and the least expressed gene a value of 2.33. These measurements would suggest that a threshold of significance, which is a function of intensity, should be used and that the threshold will vary from 1.35 for the most highly expressed genes to 2.33 for the least expressed genes. However, the measurements performed here are at best a surrogate system for estimating

error in the tumor data, i.e. the above experiments control for hybridization, filter and analysis variation but do not control for labeling and other sample-handling variation in the tumor samples. With limited tissue available for each tumor, it is not possible to perform replicate measurements on all our samples to generate similar significance curves for the actual data. Hence, we chose to use a ratio of 2.5 or more as the threshold of significance for our tumor data. We recognize that this criterion will result in the exclusion of genes that are differentially regulated at a statistically level. However, given that our goal is to develop genes that may serve as serum markers for ovarian cancer, and given the limitations of currently available assay systems for serum marker testing, a factor of 2.5 differential expression is appropriate.

3.2. Screening of 21 500 ovarian clones

An ideal array of cDNAs would contain a single copy of every gene expressed by the tissues to be compared. Since the identification of all human genes is incomplete, we chose to array randomly selected cDNAs derived from a wide spectrum of ovarian tissues including normal ovarian epithelium, early stage ovarian carcinomas, and late-stage pathologically aggressive ovarian carcinomas. We chose to array 8600 clones in form of purified plasmids from an OSE library [short-term culture of ovarian surface epithelial cells (Karlan et al., 1995)], and 3225 each from four ovarian cancer cDNA libraries from increasing malignancy, totaling 21 500 arrayed clones. We created 16 replicate sets of these arrays, each set consisting of 14 membranes of 7×12 cm holding 1536 clones. Each of the membrane sets was hybridized with a ^{32}P -labeled first-strand cDNA probe made from the RNA of an early-stage serous ovarian tumor (T007), eight late-stage serous ovarian tumors (T004, T008, T009, T010, T011, T014, T015, T016), one recurrent ovarian tumor (T012), ovarian surface epithelium (N001S), liver, placenta, bone marrow, cerebellum, and kidney. Two types of comparative experiments were carried out: (1) normal and tumor ovarian tissues were contrasted, and (2) ovarian tissues were compared against a variety of normal tissues. The first comparisons would reveal the tumor-specific cDNAs and the second the ovarian-specific cDNAs (at least with respect to the five different normal tissues). It was not our purpose to analyze early-to-late stage differences or tumor stratification as the limited number of cancerous tissues would not allow this. Our objective was to determine whether it is possible to use this technique to detect genes that are overexpressed in ovarian carcinomas relative to normal ovary and other tissues.

3.3. Differential transcript expression

Using the spot-finding and detection software developed in our laboratory, we determined the hybridization

intensities for each clone and calculated their ratios. Comparing the 10 hybridizations with ovarian tumor tissues to those with OSE, the vast majority (>93%) of the clones displayed tumor-to-OSE ratios of less than a factor 2.5, and therefore were considered unchanged; about 7% of the clones exhibited a tumor-to-OSE ratio of more than 2.5, 0.9% a ratio of greater than 5.0, and 0.5% a ratio of greater than 10.0. Thus, most transcripts were expressed at similar levels in normal and tumor tissues, a finding that has been reported in colorectal and pancreatic cancers (Zhang et al., 1997).

No clone exhibited a 2.5-fold difference in expression in more than six of the ovarian tumors relative to OSE. Given the difference in tumor stages (one was an early stage tumor, and one a recurrent late stage tumor, the rest being late-stage ovarian adenocarcinomas) and the fact that the same stages, if they represent different stratified types, do not necessarily reflect high degrees of similarity on the molecular level; given the inter- and intra-tissue heterogeneity (possible proximity of section to areas of necrosis, differences in histology and pathology between tumors and across tumor sample), we did not expect to see a particular clone exhibit high tumor-to-OSE ratios in all tumors.

Sixteen clones showed overexpression in at least six ovarian cancers, but 14 of these 16 were also expressed in at least one non-ovarian tissue. In order to obtain a reasonable number of clones with overexpression in ovarian tumors and not in non-ovary tissues, we chose clones that fulfilled the following criteria: ratios greater than 2.5 in at least five out of the 10 tumors compared to OSE, and ratios below 2.5 in bone marrow, cerebellum, kidney, liver, and placenta compared to OSE. We were able to identify 134 clones that fulfilled these criteria. Sequencing of the partial cDNA clones revealed 60 that matched sequences in the non-redundant (nr) GenBank database. Of these, 17 matched to mitochondrial and ribosomal genes, and 43 matched to 37 other characterized genes (Table 1). Forty-seven clones matched only to sequences in the EST database, and 24 clones did not match any sequence in GenBank and were classified as novel. Three clones of 254, 312 and 323 bp length matched entirely to SINE and LINE sequences and were thus classified as repeats (see Table 1).

The expression patterns of two of these clones, which code for *S-adenosyl homocysteine* hydrolase and *HE4*, are shown in Fig. 2. For both genes, the calculated overexpression by signal intensities in the cancer tissues can be confirmed by visual inspection of the hybridized membranes. It is obvious, however, that by visual inspection alone, these clones would have probably escaped our scrutiny since their expression is rather weak compared to neighboring clones.

The overexpression of the 17 clones with similarity to mitochondrial sequences and ribosomal proteins can

Table 1
Categories of cDNAs present in the 134 clones^a

Number of sequences	Percentage	Sequence similarity
3	2	Repeats
6	4	Mitochondrial sequences
2	2	Ribosomal RNA
9	7	Ribosomal proteins
24	18	Novel sequences
47	35	ESTs (expressed sequence tags)
43	32	Known genes
134		Total

^a Novel sequences had less than 60% similarity to either human or non-human sequences. Repeats: genomic, SINE (ALU, MIR) and LINE (LINE1 and LINE2), LTR elements (MaLRs, Retroviral, MER4 group), DNA elements (MER1, MER2, Mariner). GenBank Accession Nos of the clones with similarity to known genes: 14.3.3, X56468 (2×); *Actin capping protein*, U03269; *alpha-enolase*, M14328; *beta-actin*, M10277; *beta-2 microglobulin*, M17987; *BA46*, U58516; *Catechol-O-methyltransferase*, M65212; *CD44*, L05412; *CLIP/Restin*, M97501/X64838; *E16*, M80244; *Elongation factor 1 beta*, X60489; *Elongation factor 1 gamma*, Z11531 (2×); *Elongation factor 2*, Z11692; *Flightless*, U01184; *HE4*, X63187 (2×); *Initiation factor 4AI*, D13748; *Insulin-like growth factor BP 3 precursor*, M31159; *MDC15*, U46005; *Mucin*, X52229; *Myosin*, M22918; *Oviductal glycoprotein*, U09550 (3×); *p84*, L36529; *Peroxisomal targeting signal receptor 1*, U19721; *Phosphatidyl inositol-3-kinase alpha subunit*, M61906; *Poly-A binding protein*, Y00345; *Procollagen alpha COL1A2*, K01078; *putative Progesterone binding protein*, Y12711; *Proteasome subunit HC8*, D00762; *RhoA*, L25080; *Ryudocan*, D13292; *S-adenosyl-homocysteine hydrolase*, M61831; *Smooth muscle protein*, M95787; *Tenascin precursor*, X56160; *Thiol-specific antioxidant*, Z22548; *Thymosine beta 4* (interferon-inducible), M17733; *Tropomyosin*, M75165; *Ubiquitin*, M10939, X56997 (2×).

be attributed to the higher metabolic activity of the tumors. Ribosomal protein sequences have been found to be more highly expressed in colon carcinomas (Pogue-Geile et al., 1991). Likewise, five other genes linked to metabolic pathways such as *elongation factor 1 gamma* and *initiation factor 4AI* were overexpressed in ovarian cancer tissues. It is notable that these 22 clones displayed an average tumor-to-OSE ratio of 5.22 ± 2.4 , whereas the remaining 38 clones with homology to known genes had a lower average ratio of 4.11 ± 1.8 . This underscores the fact that the degree of overexpression alone is not necessarily indicative of a clone that can be used as a marker protein.

In order to estimate the quality of the HDAH in identifying cancer related genes, and since we were realistically capable of processing only a limited number of clones, we focused on the 43 previously characterized clones, as opposed to the 47 clones that match only ESTs or those 24 that do not match any sequence in GenBank. Of the 43 clones with homology to the 37 characterized genes, 10 genes are expressed in epithelial tissues: 14.3.3, *BA46*, *CD44*, *HE4*, *Mucin1*, *Oviductal glycoprotein*, *Collagen COL1A2*, *Putative progesterone binding protein*, *RhoA*, and *Ryudocan* (GenBank Accession Nos listed in Table 1). This coincides with

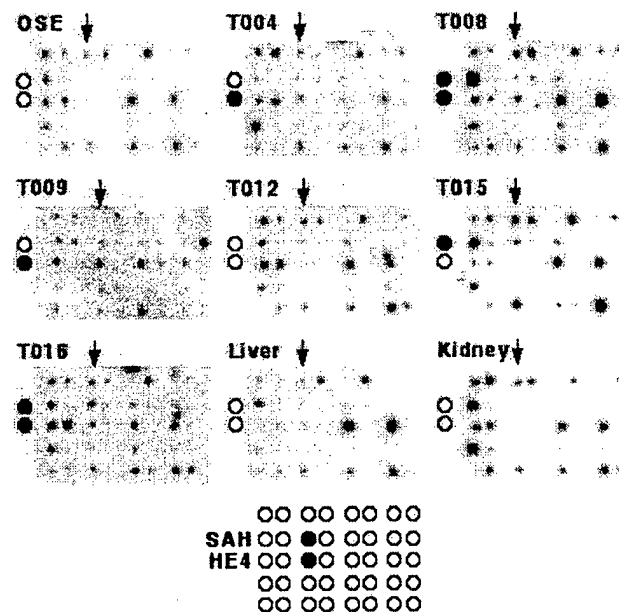


Fig. 2. Visual inspection of the membranes after identification of differentially expressed clones through computing the signal intensities. Displayed are eight columns and five rows of close-ups on nine membranes that have been hybridized with nine different probes. Two clones, coding for *HE4* and *S-adenosyl homocysteine hydrolase (SAH)*, show high tumor-to-OSE ratios and low normal-to-OSE ratios (see Table 2). The positions of these two clones on the array are marked in the bottom panel. The clone positions on the nine membranes are indicated by an arrow on top of each panel and the circles on the left. An empty circle denotes a weak hybridization signal, and a filled circle denotes a strong signal.

the fact that the vast majority of ovarian cancers, including all those used for HDAH, arise from the ovarian surface epithelium (Berchuck et al., 1996).

Thirteen of the 37 genes (35%) are known to be overexpressed in various cancers, including lung, breast and colon. Six of these 13 are expressed in ovarian carcinomas, their expression not being restricted to ovarian tissues. The thirteen genes are 14.3.3 [lung cancer (Nakanishi et al., 1997)], *beta-actin* [AML (Blomberg et al., 1987) and colorectal carcinomas (Naylor et al., 1992)], *BA46* [breast cancer (Couto et al., 1996)], *CD44* [ovarian cancer cell lines (Stickeler et al., 1997)], *Clip/Restin* [Hodgkin disease and anaplastic large-cell lymphoma (Delabie et al., 1992)], *Collagen COL1A2* [ovarian cystadenoma (Kauppila et al., 1996)], *E16* [colorectal carcinoma, adenocarcinomas from breast and endometrium (Wolf et al., 1996)], *Insulin-like growth factor BP 3* [breast cancer (Ng et al., 1998)], *Mucin1* [epithelial ovarian cancer (Dong et al., 1997)], *Procollagen-alpha* [ovarian cystadenocarcinoma (Kauppila et al., 1996)], *putative Progesterone binding protein* [ovarian cancer (Isola et al., 1990)], *RhoA proto oncogene* [*ras* activation (Khosravi-Far et al., 1995)], and *MDC15*, a metalloprotease [some metalloproteases are elevated in ovarian tumor cell cultures (Fishman et al., 1997)]. These findings indicate that our approach

is indeed capable of narrowing down the pool of 21 500 randomly selected clones to a few epithelium- and cancer-related genes.

3.4. Confirmation of overexpression of four selected clones by RT-PCR-based transcript quantitation

Any clone with its expression restricted to ovarian carcinomas can be potentially used as a marker without knowing its function. Early detection of ovarian cancer, however, requires that the assay be suitable for routine screening of women, which means that it must be affordable, non-invasive and with a high degree of specificity. Only a serum-based assay can deliver this. Therefore, knowing whether a protein is secreted or membrane-bound maximizes the chance that the protein or its degradation product will be found in the blood either as freely circulating protein or bound to the membrane of a cell that has detached from the tumor. In both cases, an antibody can be used to detect the protein in the blood. A circulating cancer cell can be detected by an RT-PCR assay or fluorescence-activated cell sorting.

In an attempt to find out whether one of the 43 clones that match characterized genes would be a potential candidate for a marker protein in a serum-based assay, we examined which of the clones codes for a cell surface protein such as Her2/neu, used as a target in breast cancer treatment (Baselga et al., 1998) or a secreted protein such as Prostate Specific Antigen (PSA) which is used in prostate cancer diagnosis (Rittenhouse et al., 1998).

From the 43 clones with homology to the 37 known genes, we chose five that are expressed at the cell surface (*progesterone binding protein*, *ryudocan*, *mucin1*, *E16*, *BA46*) and one which is secreted (*HE4*). In addition, we included the gene *14.3.3*, which is expressed in the cytoplasm but which, like *HE4*, appeared twice in our selected clones list. *Beta actin* is often used as a control for quantitative analyses because of its assumed uniformity in expression in a large array of tissues. Our HDAH results suggest, however, that *beta actin* is differentially expressed in some ovarian tumors. We therefore chose to verify *beta actin* expression as well. The characteristics of the eight chosen genes are summarized in Table 2. We used RT-PCR-based transcript quantitation to confirm overexpression in tumors relative to normal tissues.

Due to the small size of our tumor specimens (ranging from 200 to 400 mg per tissue), the RNA preparations used in the array hybridization were exhausted during library construction and probe preparation. Therefore, new ovarian adenocarcinomas matching the stage and grade of the original tumors were used for the RT-PCR analysis. We chose one early-stage, low-grade mucinous ovarian adenocarcinoma (T037) five late-stage, high-grade serous ovarian adenocarcinomas (T001–T006 and

T021) and two metastatic ovarian serous adenocarcinomas (T013M and T026M). In order to incorporate different tumor histologies, we included two benign ovarian tissues (T017B and T018B) as well as a borderline ovarian tumor tissue (T028L). In addition, we tested the expression in four normal ovaries (N002, N005, N006 and N019), in a pool of fetal ovaries and in two batches of peripheral blood lymphocytes (PBL1 and PBL2). The reason for analyzing the expression patterns of these genes in peripheral blood lymphocytes is to determine whether they are expressed in blood elements, for if they are, they would not be good candidates for a diagnostic probe in blood samples. The OSE, as well as the liver and placental tissue were the same as used for array hybridization. As a control for the quality of the RNA template, we included a gene that we found to be expressed at high levels in all tissues tested so far, *S31iii125* (GenBank Accession No. U61734, Trower et al., 1996).

Fig. 3 shows the results of the RT-PCR. The quantitated intensities of the PCR bands are summarized in Table 2. While trying to match the tumor tissues in stage and grade, we did not expect an exact reproduction of the ratios from the HDAH analysis. In spite of these shortcomings, we were able to reproduce the tumor-to-OSE ratios observed in the HDAH for seven out of the eight genes, albeit only qualitatively. For the gene *14.3.3*, the tumor-to-OSE ratios were low but still measurable. This discrepancy can be attributed to the difference in tumor samples used or to an erroneous reading of the HDAH signals. For three genes (*BA46*, *E16* and *Ryudocan*), a high placenta-to-OSE ratio stands in discordance with the HDAH results where they had been low. Since the placental RNA used in both cases was the same, and since our quadruple RT-PCR approach is more accurate than the HDAH method, we must conclude that in the HDAH, the placental values must have been misread for these three clones.

14.3.3 shows no tumor-to-OSE ratios above the threshold of significance of 2.5. It displays a mean ratio of 1.5 in four invasive and in one benign ovarian tumor, which does not compare well with the mean ratio of 4.4 determined in the HDAH.

BA46 shows tumor-to-OSE ratios above 2.5 in five tumors but also in one normal ovary and in placenta. In spite of its low expression in PBL (which, as noted in the beginning of this section, is a prerequisite for a serum marker), the relatively low mean ratios in RT-PCR and HDAH of 3.2 make it a second choice marker gene.

Beta actin shows tumor-to-OSE ratios above 2.5 in 10 out of the 12 tumors (a mean of 3.9 compared to 4.4 in the HDAH), but also in some normal tissues, including PBL. Although these numbers do not warrant the consideration as a tumor marker gene, they give cause to question the use of *beta actin* as a normalization standard.

Table 2
HDAH (top) and RT-PCR ratios (bottom) of nine selected genes*

Gene name	14.3.3	14.3.3	BA46	β -actin	E16	HE4	HE4	Mucin1	ProgBP	Ryu
Accession No.	X56468	X56468	U58516	X00351	M80244	X63187	X63187	X52229	Y12711	D13292
Protein	Cytopl.	Cytopl.	Membr	Cytopl.	Membr	Secreted	Secreted	Membr.	Membr	Membr
T004				3.1		5.1	3.6		8.9	3.1
T007 early	5.9	3.7			5.5	3.0	2.7	2.6		
T008			4.1			5.1	4.9	2.6		5.0
T009	2.7			8.5		5.1	5.5		2.8	
T010	5.5	4.3	2.7	2.5						
T011			2.6		3.0		2.7			
T012 recur						2.8		8.0	3.6	
T014	6.8	3.0		4.2	5.2			5.9	7.3	
T015	6.0	2.7	2.7		4.5				2.9	4.1
T016		3.1	4.0	3.6	3.7	2.5	2.6	8.4		4.6
Liver				3.8						2.5
Placenta			2.5	3.4	9.8				1.3	2.1
PBL1	1.3	1.3		3.9					5.8	2.1
PBL2				4.7	3.9			2.0	4.1	2.3
Fetal	2.8	2.8	1.9	1.3	8.9	7.9	7.9	9.7	4.9	1.9
N002			2.1	3.8	2.4			3.7	1.8	1.4
N005			1.3	3.4	6.5	1.2	1.2	2.4	2.1	1.0
N006			2.6	3.5	4.5	2.0	2.0	2.3	4.1	0.6
N019									3.2	0.3
N035			3.1	4.2				2.1		1.3
T017B	1.6	1.6	1.6	2.7	2.0	6.6	6.6	2.3	2.6	5.3
T018B			2.5	3.6	3.3	8.8	8.8	2.8	7.8	
T028L				2.9	1.1	8.2	8.2	3.2	1.3	4.9
T037 early			1.9	3.2	1.4	1.6	1.6	7.9	3.1	4.4
T002			1.0	3.0	1.6	12.0	12.0	2.4	2.7	3.3
T003	1.6	1.6	3.7	3.1	1.5	16.0	16.0	1.9	4.7	3.7
T005			3.0	3.6	9.4	17.0	17.0	2.4	2.0	4.8
T006				2.5		9.7	9.7		2.8	2.1
T001	1.3	1.3	1.1	2.4	2.2	11.4	11.4	2.4		2.7
T021			3.0	5.7	1.2	12.3	12.3	2.5	2.1	3.7
T013M	1.9	1.9		5.2	3.4	14.1	14.1	7.3	9.7	1.4
T026M	1.2	1.2	4.0	5.8	1.9	2.8	2.8	4.5	3.5	

* Eight genes out of the 43 clones that match to 37 known genes were validated for their expression by RT-PCR (see Fig. 3). The volumes of the PCR bands were calculated using the software QuantityOne (BioRad, Hercules, CA). Titan RT-PCR amplifies the template semiquantitatively; therefore, the numbers in this table are merely indicative of a tendency and cannot be translated into copy numbers. The rows show the gene name, GenBank Accession No., protein localization, 10 tumor-to-OSE ratios that were observed in the HDAH (only ratios above 2.5; normal-to-OSE ratios are omitted for they lied all below 2.5), followed by 22 tissue-to-OSE ratios determined in the RT-PCR (for clarity, only ratios above 1 are displayed). The columns are duplicated for 14.3.3. and HE4 because two clones were selected for them by HDAH. *Putative Progesterone binding protein* (ProgBP): progesterone binding proteins can be found in low-grade breast cancers and in some ovarian cancer cell lines. The homologous rat sequence has a transmembrane region (Falkenstein et al., 1996), indicating that our clone might also be membrane-bound. *Ryudocan* (abbreviated as Ryu.) is a cell-surface proteoglycan with a transmembrane domain; it is expressed in an extensive array of human tissues (Kojima et al., 1993). *HE4* is an epidermal, epididymis-specific protease inhibitor that is thought to be involved in the maturation of spermatozoa (Kirchhoff et al., 1991). The putative *HE4* protein has a leader sequence and it is speculated that it is secreted. *Mucin1* (Dong et al., 1997) is expressed on the cell surface of non-mucinous ovarian tumors with either low malignant or invasive potential. 14.3.3 codes for a cytosolic protein kinase regulator protein that shows elevated expression levels in lung cancer tissues (Nakanishi et al., 1997). *BA46*, also known as lactadherin, is a cell-surface protein expressed in human breast carcinomas. It has been used successfully as a target for experimental breast cancer radioimmunotherapy (Couto et al., 1996). *Beta actin* is a cytoskeletal protein with differential expression in acute myelolytic leukemia (Blomberg et al., 1987) and high expression in colorectal carcinomas (Naylor et al., 1992). *E16* codes for an integral membrane protein that was isolated from peripheral blood lymphocytes (Gaugitsch et al., 1992). It is expressed in colorectal and other human carcinomas (Wolf et al., 1996).

E16 shows tumor-to-OSE ratios above 2.5 in three tumors (with a mean of 5.3 compared to a mean of 4.4 in the HDAH). It also shows high ratios for two normal ovaries and placenta. The low expression in PBL and the high average ratios for the tumors make it a possible marker candidate.

HE4 shows a clear tumor-restricted expression,

making its pattern resemble that in the HDAH. Most importantly, the results suggest that it is not expressed in peripheral blood lymphocytes. As noted in the beginning of this section, this accordingly represents a candidate for a serum marker assay. The difference in the mean rates of overexpression measured by RT-PCR (11×) and HDAH (4.1×) can be attributed either to

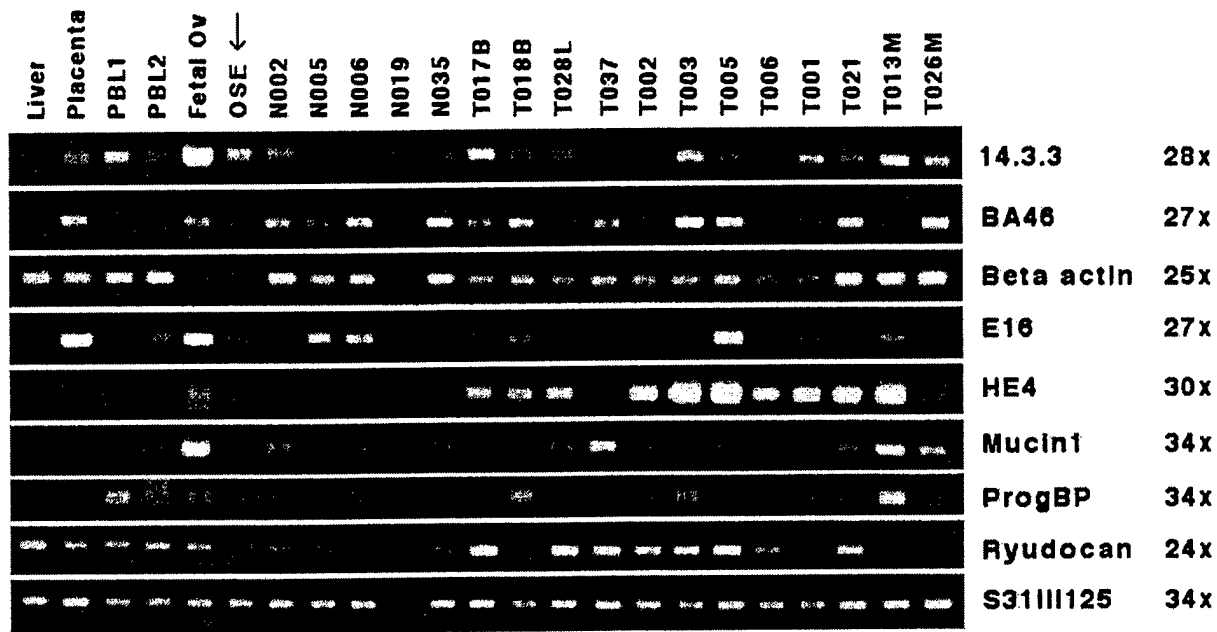


Fig. 3. Expression monitoring by RT-PCR. Eight genes (plus one control) are tested in 23 tissues. Tissue names are on top; OSE is marked with an arrow. Tissues starting with an N are normal ovaries, and those starting with a T are ovarian tumors. The *S31III125* gene serves as a control. The number of PCR cycles is indicated behind each gene name (on the right). ProgBP stands for 'putative progesterone binding protein'. The PCR bands are in the range of 420–660 bp. All reactions had been performed in four parallel sets with one set shown here.

the better signal-to-noise ratio in the RT-PCR or to the different tumor samples used.

Mucin 1 shows a high RT-PCR value in fetal ovaries, suggesting that this might be a fetal gene that is re-expressed in the tumor. It shows strong bands in three out of the 12 tumors, two of them metastatic and one an early stage tumor, resulting in a mean tumor-to-OSE ratio of 3.0. This result correlates with that of the array hybridization.

The *Putative progesterone binding protein* shows a high tumor-to-OSE ratio for only two tumors, one being similar in stage to a tumor used in the HDAH. All other tumors show medium high ratios but so do the normal tissues, including the PBL. The strong expression in the metastasizing tumor may indicate a role as a marker for tumor staging, prognosis or stratification.

The transcript of *ryudocan* displays a similar pattern of expression as *HE4*, and the mean the tumor-to-OSE ratio of 4.3 are is slightly higher than the one determined by HDAH (where it was 6). The presence of *ryudocan* mRNA in liver, PBL and placenta means that the protein might normally be found in the blood, thus making it a less suitable marker candidate.

3.5. Confirmation of overexpression of *HE4* by Northern blot analysis

Of the eight genes tested in the RT-PCR, only *HE4* shows a clear tumor-restricted expression pattern. To further confirm the cancer-restricted expression of *HE4*, we used a Northern blot (Northern Territory[®],

Invitrogen, San Diego, CA) that contained total RNA from ovaries from four patients who had unilateral ovarian cancer. RNA from both the affected and the unaffected ovary was present on the blot (loaded adjacent to each other). Fig. 4 shows that *HE4* is expressed in two ovarian carcinomas but not in the matching normal ovaries. *HE4* cannot be detected in the tumors nor in the normal ovaries of two other patients. The ratios of *HE4* expression between the unaffected and

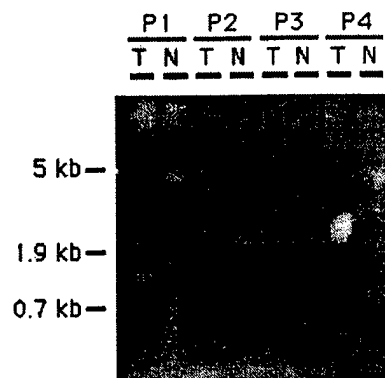


Fig. 4. Northern hybridization of *HE4*. The Northern blot contains RNA from ovarian tumor and matching non-affected ovary from four patients. *HE4* is expressed in two tumors but not in the normal tissue of the same patient. A Digoxigenin-labeled riboprobe was prepared from a 500 bp *HE4* PCR product cloned in a vector. The probe was hybridized over night at 68°C and washed in 2 × SSC, 0.1% SDS for 15 min at room temperature; 2 × SSC, 0.1% SDS for 20 min at 68°C; 0.1 × SSC, 0.1% SDS for 2 × 15 min at 68°C. The hybridized RNA was visualized using the DIG detection kit (Boehringer Mannheim). The membrane was exposed to X-ray film for 15 min.

the affected ovary was 6.1 for patient 1 and 4.5 for patient 2. Thus, *HE4* is also a candidate for a tumor-staging, prognosis or stratification marker.

3.6. Conclusion

From the 21 500 clones, we chose 43 that were overexpressed in ovarian tumors by HDAH with homology to characterized genes. We chose eight genes for expression validation by RT-PCR. From these eight, seven genes displayed tumor-to-OSE ratios similar to those measured in the HDAH, albeit with different tumor tissues matching grade and stage. Seven of these eight display expression in normal tissues; only *HE4* showed a clear tumor-restricted expression pattern. We conclude that the *HE4* message is significantly overexpressed in a variety of ovarian tumors relative to normal tissues or OSE, thus making it a potential candidate for a marker protein.

The results support the validity of using HDAH combined with a second quantitation method for the identification of genes that are overexpressed in cancers as compared to normal tissues. We are preparing an antibody against *HE4* to further analyze whether it indeed could be a diagnostic marker for ovarian cancer.

Acknowledgements

This work was supported by the Stowers Institute and by grants from the Deutsche Forschungsgemeinschaft, the National Institutes of Health (5RO1HG01713-02) the Marsha Rivkin Center for Ovarian Cancer Research and the National Science Foundation (BIR9214821/9423347). We would like to thank the members of the sequencing group in our laboratory.

References

- American Cancer Society, 1998. Cancer Facts and Figures. American Cancer Society, Atlanta, GA.
- Baselga, J., Norton, L., Albanell, J., Kim, Y.M., Mendelsohn, J., 1998. Recombinant humanized anti-HER2 antibody (Herceptin) enhances the antitumor activity of paclitaxel and doxorubicin against HER2/neu overexpressing human breast cancer xenografts. *Cancer Res.* 58, 2825–2831.
- Berchuck, A., Kohler, M.F., Bast Jr., R.C., 1996. Molecular genetic features of ovarian cancer. *Prog. Clin. Biol. Res.* 394, 269–284.
- Bishop, J.O., Morton, J.G., Rosbash, M., Richardson, M., 1974. Three abundance classes in HeLa cell messenger RNA. *Nature* 250, 199–204.
- Blomberg, J., Andersson, M., Faldt, R., 1987. Differential pattern of oncogene and beta-actin expression in leukaemic cells from AML patients. *Br. J. Haematol.* 65, 83–86.
- Couto, J.R., Taylor, M.R., Godwin, S.G., Ceriani, R.L., Peterson, J.A., 1996. Cloning and sequence analysis of human breast epithelial antigen BA46 reveals an RGD cell adhesion sequence presented on an epidermal growth factor-like domain. *DNA Cell Biol.* 15, 281–286.
- Delabie, J., Shipman, R., Brüggen, J., De Strooper, B., van Leuven, F., Tarcsay, L., Cerletti, N., Odink, K., Diehl, V., Bilbe, G., et al., 1992. Expression of the novel intermediate filament-associated protein restin in Hodgkin's disease and anaplastic large-cell lymphoma. *Blood* 80, 2891–2896.
- Dong, Y., Walsh, M.D., Cummings, M.C., Wright, R.G., Khoo, S.K., Parsons, P.G., McGuckin, M.A., 1997. Expression of MUC1 and MUC2 mucins in epithelial ovarian tumours. *J. Pathol.* 183, 311–317.
- Falkenstein, E., Meyer, C., Eisen, C., Scriba, P.C., Wehling, M., 1996. Full-length cDNA sequence of a progesterone membrane-binding protein from porcine vascular smooth muscle cells. *Biochem. Biophys. Res. Commun.* 229, 86–89.
- Fishman, D.A., Bafetti, L.M., Banionis, S., Kearns, A.S., Chilukuri, K., Stack, M.S., 1997. Production of extracellular matrix-degrading proteinases by primary cultures of human epithelial ovarian carcinoma cells. *Cancer* 80, 1457–1463.
- Gaugitsch, H.W., Prieschl, E.E., Kalthoff, F., Huber, N.E., Baumrucker, T., 1992. A novel transiently expressed, integral membrane protein linked to cell activation. Molecular cloning via the rapid degradation signal AUUUA. *J. Biol. Chem.* 267, 11267–11273.
- Isola, J., Kallioniemi, O.P., Korte, J.M., Wahlstrom, T., Aine, R., Helle, M., Helin, H., 1990. Steroid receptors and Ki-67 reactivity in ovarian cancer and in normal ovary: correlation with DNA flow cytometry, biochemical receptor assay and patient survival. *J. Pathol.* 162, 295–301.
- Jacobs, I., Davies, A.P., Bridges, J., Stabile, I., Fay, T., Lower, A., Grudzinkas, J.G., Oram, D., 1993. Prevalence screening for ovarian cancer in postmenopausal women by CA 125 measurement and ultrasonography [see comments]. *Br. Med. J.* 306, 1030–1034.
- Karlan, B.Y., Jones, J., Greenwald, M., Lagasse, L.D., 1995. Steroid hormone effects on the proliferation of human ovarian surface epithelium in vitro. *Am. J. Obstet. Gynecol.* 173, 97–104.
- Kaupilla, S., Saarela, J., Stenback, F., Risteli, J., Kaupilla, A., Risteli, L., 1996. Expression of mRNAs for type I and type III procollagens in serous ovarian cystadenomas and cystadenocarcinomas. *Am. J. Pathol.* 148, 539–548.
- Khosravi-Far, R., Solski, P.A., Clark, G.J., Kinch, M.S., Der, C.J., 1995. Activation of Rac1, RhoA, and mitogen-activated protein kinases is required for Ras transformation. *Mol. Cell. Biol.* 15, 6443–6453.
- Kirchhoff, C., Habben, I., Ivell, R., Krull, N., 1991. A major human epididymis-specific cDNA encodes a protein with sequence homology to extracellular proteinase inhibitors. *Biol. Reprod.* 45, 350–357.
- Kojima, T., Inazawa, J., Takamatsu, J., Rosenberg, R.D., Saito, H., 1993. Human ryudocan core protein: molecular cloning and characterization of the cDNA, and chromosomal localization of the gene. *Biochem. Biophys. Res. Commun.* 190, 814–822.
- Nakanishi, K., Hashizume, S., Kato, M., Honjoh, T., Setoguchi, Y., Yasumoto, K., 1997. Elevated expression levels of the 14-3-3 family of proteins in lung cancer tissues. *Hum. Antibodies* 8, 189–194.
- Naylor, M.S., Stamp, G.W., Balkwill, F.R., 1992. Beta actin expression and organization of actin filaments in colorectal neoplasia. *Epithelial Cell Biol.* 1, 99–104.
- Ng, E.H., Ji, C.Y., Tan, P.H., Lin, V., Soo, K.C., Lee, K.O., 1998. Altered serum levels of insulin-like growth-factor binding proteins in breast cancer patients [in process citation]. *Ann. Surg. Oncol.* 5, 194–201.
- Ng, W.L., Schummer, M., Cirisano, F., Baldwin, R.L., Karlan, B.Y., Hood, L., 1996. High-throughput plasmid miniprepations facilitated by micro-mixing. *Nucleic Acids Res.* 24, 5045–5047.
- Piétu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E.,

- Mariage-Samson, R.R.H., Soularue, P., Auffray, C., 1996. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.* 6, 492–503.
- Pogue-Geile, K., Geiser, J.R., Shu, M., Miller, C., Wool, I.G., Meisler, A.I., Pipas, J.M., 1991. Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. *Mol. Cell. Biol.* 11, 3842–3849.
- Rittenhouse, H.G., Finlay, J.A., Mikolajczyk, S.D., Partin, A.W., 1998. Human Kallikrein 2 (hK2) and prostate-specific antigen (PSA): two closely related, but distinct, kallikreins in the prostate [in process citation]. *Crit. Rev. Clin. Lab. Sci.* 35, 275–368.
- Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E., Davis, R.W., 1998. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 16, 301–306.
- Schummer, B., Hauptfleisch, S., Siegmund, M., Schummer, M., Lemmer, B., 1998. Highly accurate quantification of mRNA expression by means of Titan[®] One Tube RT-PCR and capillary electrophoresis. *Biochemica* 2, 31–33.
- Schummer, M., Ng, W.-L., Nelson, P.S., Bumgarner, R.B., Hood, L., 1997. A simple high-performance DNA arraying device for comparative expression analysis of a large number of genes. *BioTechniques* 23, 1087–1092.
- Stickeler, E., Runnebaum, I.B., Möbus, V.J., Kieback, D.G., Kreienberg, R., 1997. Expression of CD44 standard and variant isoforms v5, v6 and v7 in human ovarian cancer cell lines. *Anticancer Res.* 17, 1871–1876.
- Trower, M.K., Orton, S.M., Purvis, I.J., Sanseau, P., Riley, J., Christodoulou, C., Burt, D., See, C.G., Elgar, G., Sherrington, R., Rogaev, E.I., St George-Hyslop, P., Brenner, S., Dykes, C.W., 1996. Conservation of synteny between the genome of the pufferfish (*Fugu rubripes*) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease. *Proc. Natl. Acad. Sci. USA* 93, 1366–1369.
- Wolf, D.A., Wang, S., Panzica, M.A., Bassily, N.H., Thompson, N.L., 1996. Expression of a highly conserved oncofetal gene, TA1/E16, in human colon carcinoma and other primary cancers: homology to *Schistosoma mansoni* amino acid permease and *Caenorhabditis elegans* gene products. *Cancer Res.* 56, 5012–5022.
- Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B., Kinzler, K.W., 1997. Gene expression profiles in normal and cancer cells. *Science* 276, 1268–1272.

Tissue Classification with Gene Expression Profiles

Amir Ben-Dor *
U. Washington

Laurakay Bruhn
HP Laboratories

Nir Friedman
Hebrew University

Iftach Nachman
Hebrew University

Michèl Schummer
U. Washington

Zohar Yakhini
HP Laboratories

September 30, 1999

Abstract

Constantly improving gene expression profiling technologies are expected to provide understanding and insight into cancer related cellular processes. Gene expression data will also significantly aid in the development of efficient cancer diagnosis and classification platforms. In this work we examine two sets of gene expression data measured across sets of tumor and normal clinical samples. One set consists of 2,000 genes, measured in 62 epithelial colon samples [1]. The second consists of $\approx 100,000$ clones, measured in 32 ovarian samples [24, 25].

We examine the use of scoring methods, measuring separation of tumors from normals using individual gene expression levels. These are then coupled with high dimensional classification methods to assess the classification power of complete expression profiles. We present results of performing *leave-one-out cross validation* (LOOCV) experiments on the two data sets, employing SVM [6], AdaBoost [12] and novel clustering based classification techniques. As tumor samples can differ from normal samples in their cell-type composition we also perform LOOCV experiments using appropriately modified sets of genes, eliminating the resulting bias.

We demonstrate success rate of at least 90% in tumor vs normal classification, using sets of selected genes, with as well as without cellular contamination related members. These results are insensitive to the exact selection mechanism, over a certain range.

*Contact author. Email: amirbd@cs.washington.edu.

1 Introduction

The process by which the approximately 100,000 genes encoded by the human genome are expressed as proteins involves two steps. First, DNA sequences are transcribed into mRNA sequences which in turn are translated into the amino acid sequences of the proteins that perform various cellular functions. A crucial aspect of proper cell function is that the gene expression process is regulated such that different cell types express different subsets of genes. Measuring mRNA levels can provide a detailed molecular view of the subset of genes expressed in different cell types. Recently, array-based methods have been developed that enable simultaneous measurements of the expression levels of thousands of genes. These measurements are made by quantitating the hybridization (detected for example, by fluorescence) of a cellular mRNA mixture to an array of defined cDNA or oligonucleotide sequences immobilized on a solid substrate. Array methodologies have led to a tremendous acceleration in the rate at which gene expression pattern information is accumulated [14, 17, 7, 28, 15]. Measuring gene expression levels under different conditions is important for expanding our understanding of gene function, how various gene products interact, and how experimental treatments can affect cellular function.

One of the promising usages of gene expression measurements, is the understanding of cancer. Normal cells can evolve into malignant cancer cells through a series of mutations in genes that control the cell cycle, apoptosis, and genome integrity, to name only a few. As determination of cancer type and stage is often crucial to the assignment of appropriate treatment [10], a central goal is the identification of sets of genes that can serve, via expression profiling assays, as classification or diagnosis platforms.

Another important application of these tools, is the understanding of cellular responses to drug treatment. Expression profiling assays performed before, during and after treatment, are aimed at identifying drug responsive genes, indications of treatment outcomes, and at identifying potential drug targets [5]. More generally, complete profiles can be considered as a potential basis for classification of treatment progression or other trends in the evolution of the treated cells.

Data obtained from such studies typically consists of expression level measurements of thousands of genes. This complexity calls for data analysis methodologies that will efficiently aid in extracting relevant biological information. Previous gene expression analysis work emphasizes clustering techniques, which aim at partitioning the set of genes into subsets that are expressed similarly across different conditions. Indeed, such clustering has been demonstrated to identify functionally related families of genes [2, 7, 4, 13, 28, 9]. Similarly, clustering methods can be used to divide a set of cell samples into clusters based on their expression profile. In [1] this approach was applied, and a set of colon samples was divided into two groups, one containing mostly tumor samples, and the other containing mostly normal tissue samples.

Clustering methods, however, do not use any tissue annotation (e.g., tumor vs. normal) in the partitioning step. This information is used only afterward, to assess the success of the method. Such methods are often referred to as *unsupervised*. In contrast, *supervised* methods, attempt to predict the classification of new tissues, based on their gene expression profiles after training on examples that have been classified by an external "supervisor".

The purpose of this work is to rigorously assess the potential of classification approaches on gene expression data. We present a novel clustering based classification methodology, and apply it together with two other recently developed classification approaches, *Boosting* and *Support Vector Machines* to two data sets. Both sets involve corresponding tissue samples from tumor and normal biopsies. The first is the data set of colon cancer [1], and the other is a data set of ovarian cancer [24]. We use established statistical tools to evaluate the predictive power of these methods in the data sets. For this purpose we use *leave one out cross validation* (LOOCV), a well known method for estimating classification accuracy.

One of the major challenges of gene expression data is the large number of genes in the data sets. For example, one of our data sets includes over 97,800 clones. Many of these clones are not relevant to the distinction between cancer and tumor and introduce noise in the classification process. Moreover, for diagnostic purposes it is important to find small sets of genes that are sufficiently informative to distinguish between tumors and normal cells. To this end we suggest a simple combinatorial error rate score for each gene, and use this method to select informative genes. As we show, selecting relatively small subsets of genes can drastically improve the performance. Moreover, this selection process also isolates genes that are potentially intimately related to the tumor makeup.

A major challenge in a realistic assessment of the performance of such methods, is *sample contamination*. Tumor and normal samples may dramatically differ in terms of their cell-type composition. For example, in the colon cancer data [1], the authors observed that the normal colon biopsy also included smooth muscle

tissue from the colon walls. As a result, smooth muscle related genes showed high expression levels in the normal samples compared to the tumor samples. This artifact, if consistent, could contribute to success in classification. To eliminate this effect we remove the muscle specific genes and observe the effect on the success rate of the process.

Very recently, Lander et al. [10] examine gene expression profile differences in AML and ALL (two types of leukemia) biopsies. They employ scoring methods to select informative genes and perform LOOCV experiments to test voting based classification approaches.

The rest of the paper is organized as follows. In Section 2, we describe the principle classification methods we use in this study. These include two state of the art methods from machine learning, and a novel approach based on clustering algorithm of [2]. In Section 3, we describe the two data sets, the LOOCV evaluation method, and evaluate the classification methods on the two data sets. In Section 4 we address the problem of gene selection. We propose a simple method for selecting informative genes and evaluate the effect of gene selection on the classification methods. In Section 5, we examine the effect of sample contamination on possible classification. We conclude in Section 6 with a discussion of related works and future directions.

2 Classification Methods

In this section, we describe the main classification methods that we will be using in this paper. We start by formally defining the classification problem. Assume that we are given a *training set* D , consisting of pairs $\langle x_i, l_i \rangle$, for $i = 1, \dots, m$. Each *sample* x_i is a vector in \mathbf{R}^N that describes expression values of N genes/clones. The *label* l_i associated with x_i is either -1 or $+1$ (for simplicity, we will concentrate on two-label classification problems). A classification algorithm is a function f that depends on two arguments, training set D , and a query $x \in \mathbf{R}^N$, and returns a predicted label $\hat{l} = f_D(x)$. Our aim in building good classification procedures is that the predicted labels will match the “true” label of the query.

2.1 Nearest Neighbor Classifier

One of the of the simplest classification algorithms is the *nearest neighbor* classifier [8]. The intuition is simple. To classify a query x , find the most similar example in D and predict that x has the same label as that example. To carry out this algorithm we need to define a similarity measure $s(x, y)$ on expression patterns. In our experiments, we use the Pearson correlation as a measure of similarity. Formally, the classification of the nearest neighbor procedure is by the rule

$$nn_D(x) = l_i \text{ s.t. } s(x, x_i) = \max_j s(x, x_j)$$

(in situations where there are several nearest neighbors, we choose one of them arbitrarily).

This simple non-parametric classification method does not take any global properties of the training set into consideration. However, it is surprisingly effective in many types of classification problems. We use it in our analysis as a strawman, to which we compare the more sophisticated classification approaches.

2.2 Using Clustering for Classification

Recall, that clustering algorithms, when applied to expression patterns, attempt to partition the set of examples into clusters of patterns, so that all the patterns within a cluster are similar to each other, and different than patterns in other clusters. This suggests that if the labeling of patterns is correlated with the patterns, then the unsupervised clustering of the data (that does not take labels into account) would cluster patterns with the same label together and separate patterns with different labels.

Indeed, such a phenomenon is noted by Alon et al. [1] in their analysis of colon cancer. Their experiment (which we describe in more detail in Section 3), involves gene expression patterns from colon samples that include both tumors and normal tissues. They clustered patterns using a hierarchical clustering procedure (which is quite different from the one we discuss below). They note that the topmost division in the dendrogram they construct divides samples into two groups, one containing mostly tumor samples, and the other containing mostly normal tissue samples.

This suggests that for some types of classification problems, such as tumor vs. normal, clustering can distinguish among labels. Following this intuition, we build a classifier around a clustering algorithm. We first describe the clustering algorithm we use. Then, we present our clustering based classifier.

2.2.1 The clustering algorithm The BioClust algorithm [2], takes as input a threshold parameter t , which controls the granularity of the resulting clusters, and a similarity measure between the tissues¹. We say that a tissue v has *high similarity* to a set of tissues C , if the average similarity between the v and the tissues in C is at least t . Otherwise, if the average similarity is below t , we say that v has *low similarity* to C .

BioClust constructs the clusters one at a time, and halts when all tissues are assigned to clusters. Intuitively, the algorithm alternates between adding high similarity tissues to C , and removing low similarity tissues from it. Eventually, all the tissues in C have high similarity to C , while all the tissues outside of C have low similarity to C . At this stage the cluster C is closed, and a new cluster is started (See [2] for complete description of the algorithm).

Clearly, the threshold value t , has great effect on the resulting clustering. As t increases, the clusters would get smaller. At the extreme case, if t is high enough, each tissue would form a different cluster. Similarly, as t decreases, the clusters tend to get larger. If t is low enough, all tissues would be assigned to the same cluster.

2.2.2 Clustering based classifier Applying clustering algorithms for classification raises two problems. First, how do we use clustering on training data to classify a new query and, second, how do we decide which “granularity” of clustering to use? We start with the second question, and then return to the first one.

As described above, the BioClust procedure has an input parameter that determines the confidence threshold in construction of clusters. By changing this parameter, we can get different numbers of clusters and different divisions into clusters. A similar situation occurs in other clustering algorithms. For example, in hierarchical clustering algorithms (e.g., [1, 9]) we can choose different numbers of clusters by selecting a “level” of the tree. In either clustering algorithms, it is clear that attempting to partition the data to exactly two clusters, will not be the optimal choice for predicting labels. For example, if the tumor class consists of several types of tumors, then the most noticeable division into two clusters might separate “extreme” tumors from the milder ones and the normal tissues, and only further division will separate the normals from the milder tissues.

To address this question, we propose a measure of cluster *compatibility* with a given labeling. The intuition is simple: On the one hand, we want all the samples in the same cluster to have the same labels. Thus, we penalize pairs of samples that are within the same cluster but have different labels. On the other hand, we do not want to create unnecessary partitions. Thus, we also penalize pairs of samples that have the same label, but are not within the same cluster.

Formally, we define the *compatibility* score of a clusters with the training set as the sum of two terms. The first is the number of tissues pairs (v, u) such that v and u have the same label, and are assigned to the same cluster. The second term is the number of (v, u) pairs that have different labels, and are assigned to different clusters. This score is also called the *matching coefficient* in the literature [11].

It is easy to see that the two terms in this definition tradeoff the requirement that clusters should be as homogeneous as possible, and the requirement that clusters should not create small partitions. It is also important to note that we can evaluate cluster compatibility with a labeling, even when some of the patterns are not assigned a label. We simply restrict the comparison to counting pairs of examples for which we have a label.

Using this definition, we can optimize, using binary search, the choice of clustering parameters to find the most compatible clustering. That is, we consider different threshold values, t , use BioClust to cluster the tissues, and measure the compatibility of the resulting clusters with the given labels.

Finally, we choose the clustering that has maximal compatibility score to the given labeling. Thus, although the clustering algorithm is *unsupervised*, in the sense that it does not take into account the labels, we use a supervised procedure for choosing the clustering threshold. We also stress, that this general idea can be applied to other clustering methods, and is not restricted to our particular choice.

We now return to the question of prediction using clustering algorithms. To return a prediction, we examine the labels of all the patterns in the same cluster as the query. The intuition is that the query’s label should agree with the labels of most of these patterns. Thus, we can use a simple majority rule to decide on the label. If the cluster contains exactly the same number of tumor and normal tissues, then the classifier does not give a prediction for the query.

¹In this work we use the Pearson correlation between gene expression profile as the similarity measure. However, any similarity measure can be used.

2.3 Large-Margin Classifiers

The cluster-based approach we discussed in the previous section attempts to find inherent structure in the data (i.e., clusters of samples) and uses this structure for prediction. We can also use *direct* methods that attempt to learn a *decision surface* that separates the positive labeled samples from the negatively labeled samples.

The literature of supervised learning discusses a large number of methods that learn decision surfaces. These methods can be described by two aspects. First, the class of surfaces from which one is selected. This question is often closely related to the *representation* of the learned surface. Examples include linear separation (which we discuss in more detail below), decision-tree representations, and two-layer artificial neural networks. Second, the learning rule that is being used. For example, one of the simplest learning rules attempts to minimize the number of errors on the training set.

Application of direct methods in our domain can suffer from a serious problem. In gene-expression data we expect N , the number of measured genes, to be significantly larger than m , the number of samples. Thus, due to the large number of dimensions there are many simple decision surfaces that can separate the positive examples from the negative ones. This means that counting the number of training set errors is not restrictive enough to distinguish good decision surfaces from bad ones (in terms of their performance on examples not in the training set).

In this paper, we use two methods that received much recent attention in the machine learning literature. Both methods attempt to follow the intuition that classification of examples depends not only on the region they are in, but also on a notion of *margin*: how close are they to the decision surface. Classification of examples with small margins is not as confident as classification of examples with large margins. (We can think of the learned decision surface as an estimate, and thus given slightly different data we might move it a bit.) Thus, the reasoning suggests that we should select a decision surface that classifies correctly with large margin all the training examples. This basic intuition is developed in quite different manner in these two approaches. Below we discuss the intuition for both approaches, and defer additional details to the appendices.

2.3.1 Support Vector Machines *Support vector machines* (SVM) were developed in [6, 27]. A tutorial on SVMs can be found in [3]. The intuition for support vector machines is best understood in the example of linear decision rules. A linear decision rule can be represented by a hyperplane in R^N such that all examples on the one side of the hyperplane are labeled positive and all the examples on the other side are labeled negative. Of course, in sufficiently high-dimensional data we can find many linear decision rules that separate the examples. Thus, we want to find a hyperplane that is as far away as possible from all the examples. More precisely, we want to find a hyperplane that separates the positive examples from the negative ones, and also maximizes the minimum distance of the closest points to the hyperplane. This question can be posed as a quadratic program (see Appendix A), and can be solved efficiently. The resulting hyperplane can be written as weighted sum of the training examples, x_i , and the classification of a new example x can be calculated using dot products with the example vectors, $x \cdot x_i$. This treatment can be generalized to deal with training sets that are not linearly separable. We refer the reader to [3] for details.

It is clear that linear hyperplanes are a restricted form of decision surfaces. One method of learning more expressive separating surfaces is to project the training examples (and later on queries) into a higher-dimensional space, and learn a linear separator in that space. For example, if our training examples are in R^1 , we can project input values x to the vector $(1, x, x^2)$. A linear separator in the projected space is equivalent to learning an interval in the original representation of the training examples.

Thus, we can fix a projection $\Phi : R^N \mapsto R^M$ to higher dimensional space, and get more expressive decision surfaces. In this case, the classification rule for x will be composed of the inner products $\langle \Phi(x), \Phi(x_i) \rangle$. Moreover, for many projections there are *kernel* functions that compute the result of the inner product. A kernel function k for a projection Φ satisfies $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Given a legal kernel function, we can use it without knowing the actual mapping Φ .

To summarize, if we want to learn expressive decision surfaces, we can choose a kernel function, and use it instead of inner-product in the execution of the SVM optimization. This is equivalent to learning a linear hyperplane in the projected space.

In this work we consider two kernel functions:

- The linear kernel $k_1(x, y) = \langle x, y \rangle$.
- The quadratic kernel $k_2(x, y) = (\langle x, y \rangle + 1)^2$.

The rationale for using these simple kernels, is that since our input space is high dimensional, we can hope to find a simple separation rule in that space. We therefore test the linear separator, and the next order separator as a comparison to check if higher order kernels can yield better results.

2.3.2 Boosting Boosting was initially developed as a method for constructing good classifiers by repeated calls to “weak” learning procedure [20, 12]. The assumption is that we have access to a “weak learner”. Such an algorithm constructs a function $f_D(x)$ for each training set. The learner is weak in the sense that the *generalization error* of $f_D(x)$ is only slightly better than that of random guess. Formally, we assume that $f_D(x)$ classifies at least $1/2 + 1/\text{poly}(n)$ of the input space correctly.

In this paper, we use a fairly simple weak learner, that finds a simple rule of the form:

$$f(x, j, t_j, d) = \begin{cases} d & x[j] > t_j \\ -d & x[j] < t_j \end{cases}$$

where j is an index of a gene, $x[j]$ is the expression value of the j 'th gene in the vector x , t_j is a threshold corresponding to gene j , and $d \in \{+1, -1\}$ is a direction parameters. Such a classifier is called a *decision stump*. We learn decision stumps from data by exhaustively searching all genes, and for each gene search over all thresholds and directions, and finally return the combination that has the smallest number of errors.²

Boosting uses the weak learning procedure (e.g., the decision stump learner in our case) to construct a sequence of classifiers f_1, \dots, f_k , and then uses a weighted vote among these classifiers. Thus, the prediction made by the boosting algorithm has the form: $\text{sign}(\sum_j w_j f_j(x))$, where w_i are the weights assigned to the classifiers.

The crux of the algorithm is the construction of the sequence of classifiers. The intuition is simple. Suppose that we train the weak learner on the original training data D to get a classifier $f_1(x)$. Then, we can find the examples in D that are classified incorrectly by f_1 . We then want to force the learning algorithm to give these examples special attention. This is done by constructing a new training data set in which these examples are given more weight. Boosting then invokes the weak learner on the reweighted training set and obtains a new classifier. Examples are then reweighted, and the process is iterated. Thus, boosting adaptively reweights training examples to focus on the “hard” ones.³ In this paper, we use the AdaBoost algorithm of Freund and Schapire [12]. See Appendix B for the details of the algorithm. In practice boosting is an efficient learning procedure that usually has small number of errors on test sets. The theoretical understanding of this phenomenon uses a notion of margin that is quite similar to the one defined for SVMs. Recall, that boosting classification is made by averaging the “votes” of many classifiers. Define the margin of example x_i to be

$$m_i = l_i \sum_j w_j f_j(x_i).$$

By definition, we have that if $m_i > 0$, then $\text{sign}(\sum_j w_j f_j(x_i)) = l_i$, and thus x_i is classified correctly. However, if m_i is close to 0, then this classification is “barely” made. On the other hand, if m_i is close to 1, then a large majority of the classifiers make the right prediction on x_i . The analysis of Schapire et al. [18, 21] shows that the generalization error of boosting (and other voting schemes) depends on the distribution of margins of training examples. Schapire et al. also show that repeated iterations of AdaBoost continually increase the smallest margin of training examples. This is contrasted with other voting schemes that are not necessarily increasing the margin for the training set examples.

3 Evaluation

In the previous section we discussed several approaches for classification. In this section we describe empirical evaluation of the classification performance of these approaches on gene expression classifications.

²Note that for each gene, we need to consider only n rules, since the gene takes at most n different values in the training data. Thus, we can limit our attentions to mid-way points between consecutive values attained by the j 'th gene in the training data.

³More precisely, boosting distorts the distribution of the input samples. For some weak learners, like the stump classifier, this can be simulated by simply reweighting the samples.

3.1 Data sets

Before we describe the evaluation methods, we describe the two datasets we examined. Both of these data sets involve comparing tumor and normal samples of the same tissue.

Colon cancer data set. This data set is a collection of expression measurements from colon biopsy samples reported by Alon et al. [1]. The data set consists of 62 samples of colon epithelial cells. These samples were collected from colon-cancer patients. The “tumor” biopsies were collected from tumors, and the “normal” biopsies were collected from healthy parts of the colons of the same patients. The final assignments of the status of biopsy samples were made by pathological examination.

Gene expression levels in these 62 samples were measured using high density oligonucleotide microarrays. Of the ≈ 6000 genes detected in these microarray, 2000 genes were selected based on the confidence in the measured expression levels. The data set of 62 samples vs. 2000 genes is available at <http://www.molbio.princeton.edu/colondata>.

Ovarian cancer data set. This data set is a collection of expression measurements from 32 samples⁴: 15 ovary biopsies of three types of ovarian carcinomas (benign (2), mucinous (1), and serous (12)), 12 biopsies of normal ovaries, and 5 samples of other tissues (liver and blood). Gene expression levels in these 32 samples were measured using a membrane-based array with radioactive probes. The array consisted of cDNAs representing approximately 100,000 clones from ovarian clone libraries. For some of the samples, there are two or three repeated hybridizations for error assessments. In these cases, we treated the average of the reported expression levels as the expression levels in the samples.

3.2 Estimating Prediction Errors

When evaluating the prediction accuracy of the classification methods we described above, it is important not to use the *training error*. Most classification methods will perform well on examples they have seen during training. To get a realistic estimate of performance of the classifier, we must test it on examples that did not appear in the training set. Unfortunately, since we have a small number of examples, we cannot remove a portion of the examples from the training set, and use them for testing.

A common method to test accuracy in such situations is *cross-validation*. To apply this method, we partition the data into k sets, C_1, \dots, C_k of samples (typically, these will be of roughly the same size). Then, we construct a dataset $D_i = D - C_i$, that consists of all the training samples, except these in i 'th partition. We test the accuracy of the classifier $f_{D_i}()$ on samples from the partition C_i . These steps are repeated for each of the partitions. We can then estimate the accuracy of the method, by averaging the accuracy in each one of the cross-validation trials.

Cross-validation has several important properties. First, the training set and the test set in each trial are disjoint. Second, the classifier is tested on each sample exactly once. Finally, the training set for each trial is $(k - 1)/k$ of the original data set. Thus, we get a less biased estimate of the classifier behavior given a training set of size n .

There are several possible choices of k . A common approach is to set $k = n$. In this case, every trial removes one sample and trains on the rest. This method is known as *leave one out cross validation* (LOOCV). Another common choice is to set $k = 10$ or $k = 5$. LOOCV has been in use since early days of pattern recognition (e.g., [8]). In some situations, using larger partitions reduces the variance of the estimators (see [16]). In this work, we use LOOCV. However, we are in the process of collecting results using 10 fold cross validation, and will use these to reaffirm the estimates based on LOOCV in the conference version of the report.

Table 1 lists the accuracy estimates for the different methods applied to two datasets.⁵ As we can see, the clustering approach performs significantly better than the other approaches on the colon cancer data set.

3.3 ROC Curves

Estimates of classification accuracy give only a partial insight on the performance of a method. In our evaluation, we treated all errors as having equal penalty. In many applications, however, errors have asymmetric weights. To set terminology, we distinguish *false positive* errors, where normal tissues are classified as tumor,

⁴The training set contains 28 samples labeled as tumor or normal

⁵Some of our methods were not run on the ovarian cancer data set due to technical difficulties with the large number of clones. We are currently working on dealing with these technical issues.

Method	Colon	Ovarian
Nearest Neighbor	80.6% \pm 5.0%	71.4% \pm 8.5%
Clustering	88.7% \pm 4.0%	70.5% \pm 8.3%
SVM, linear kernel	80.6% \pm 5.0%	—
SVM, quad. kernel	79.0% \pm 5.1%	—
Boosting, 100 iterations	77.4% \pm 5.3%	—
Boosting, 1000 iterations	74.2% \pm 5.6%	—
Boosting, 10,000 iterations	77.4% \pm 5.3%	—

Table 1: Summary of classification accuracy of the methods on the two training sets. Reported accuracies denote average number of correct classifications and std. deviation. Estimates are based on LOOCV estimates.

^abased on the 17 predictions that the classifier made

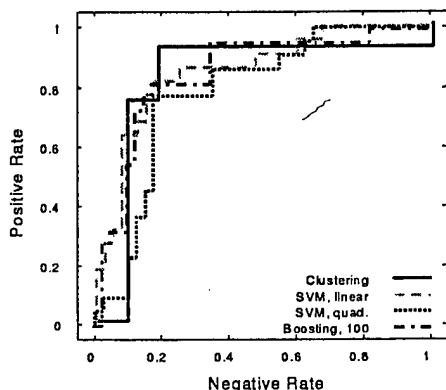


Figure 1: ROC curves for methods applied to colon cancer data set. The x -axis shows percentage of negative examples classified as positives, and y -axis shows percentage of positive examples classified as positive. Each point along the curve corresponds to the percentages achieved by a particular confidence threshold value by the corresponding classification method. Error estimates are based on LOOCV trials.

and *false negative* errors, where tumor tissues are classified as normal. In screening patients, avoiding false negative errors can be crucial, while making false positives might be tolerated (since additional tests will be performed on the patient).

To deal with asymmetric weights for errors, we introduce the *confidence parameter*, τ . In clustering approaches, the modified procedure would predict that a query tissue is tumor, if the cluster containing it has at least a fraction τ of tumors. In a similar manner, we can introduce confidence parameters for SVM and boosting approaches by changing the threshold margin needed for positive classification.

We can evaluate the “power” of a classification method for different asymmetric weights by plotting *ROC curves* (see, for example, [26]). A ROC curve plots the tradeoff between the two types of errors as we change the confidence parameters. Formally, we plot a two dimensional curve. Each point on the curve corresponds to a particular value of the confidence parameter. The (x, y) coordinates of a point specifies the fraction of negative, and positive samples that are classified as positive with this particular confidence parameters. The extreme ends of the curves are the most strict and most permissive confidence values. With the strictest confidence values, the procedure does not classify any example as positive. Thus, this value corresponds to the point $(0, 0)$. On the other hand, with the most permissive confidence value, the procedure will classify each example as positive. Thus, this confidence value corresponds to the point $(1, 1)$. The path between these two extremes shows how useful the classification method is in distinguishing between positive and negative examples. The best case scenario is that the path goes through the point $(0, 1)$. This implies that for some confidence parameter, all positives are classified as positives, and all negatives are classified as negative. The general shape of the curve and in particular the area below the curve, are indicative of the distinguishing power of a classification method.

In Figure 1 we plot the ROC curves for clustering, SVM and boosting on the colon cancer data set. As we can see, there is no clear domination among the methods. (The only exception is SVM with quadratic kernel that is consistently worse than the other methods.) The clustering procedure is clearly dominant in the region where misclassification errors are roughly of the same importance. However, SVM with linear kernel and boosting are preferred to clustering in regions of asymmetric error cost (both ends of the spectrum). We believe that the “weakness” of the clustering in the asymmetric cost regions is due to the fact that the matching coefficient score (see Section 2.2) that determines the cluster granularity treats both types of errors as having equal costs.

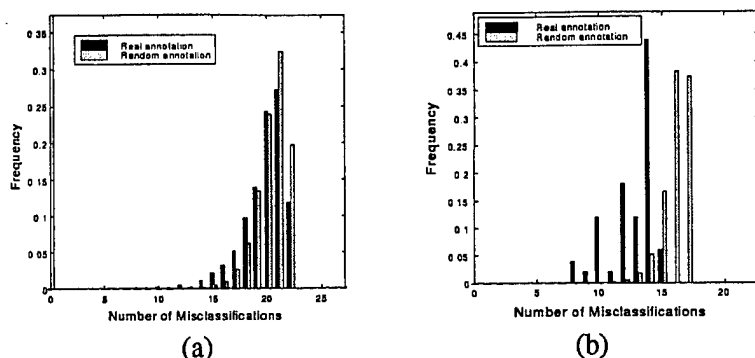


Figure 2: (a) The distribution of gene scores for the colon cancer data set comparing the scores achieved using the original labels, and the a random labeling. (b) the same histogram for the 50 best scoring genes.

4 Gene Selection

It is clear that the expression levels of many of the genes that are measured in our data sets are irrelevant to the distinction between tumor and normal tissues. Taking such genes into account during classification increases the dimensionality of the classification problem, presents computational difficulties, and introduces unnecessary noise in the process. Another issue with a large number of genes is the *interpretability* of the results. If the “signal” that allows our methods to distinguish tumor from normal tissues is encoded in the expression levels of few genes, then we might be able to understand the biological significance of these genes. Moreover, a major goal for diagnostic research is to develop diagnostic procedures based on inexpensive microarrays that have enough probes to detect diseases. Thus, it is crucial to recognize whether a small number of genes can suffice for good classification.

The problem of *feature selection* received a thorough treatment in pattern recognition and machine learning. The gene expression data sets are problematic in that they contain a large number of genes (features) and thus methods that search over subsets of features can be prohibitively expensive. Moreover, these data sets contain only a small number of samples, so the detection of irrelevant genes can suffer from statistical instabilities.

To address these issues, we propose a simple measure of “relevance” of each gene. For each gene, we measure how well we can classify the training examples if allow ourselves to ask one question about the gene’s expression level. The intuition is that an informative gene has quite different values in the two classes, and thus we should be able to separate these by a threshold value. Formally, this is equivalent to finding the best decision stump for that gene (as defined in Section 2.3.2), and then measuring how many classification errors this decision stump makes on the training examples. We call this quantity the *error score* of a gene.

An immediate question to ask is whether genes with low error scores are indeed indicative of the classification of expression. In other words, we want to test the significance of the scores of the best scoring genes in our data set. We can measure significance by analyzing the expected behavior of scores if the labeling of samples was independent of gene expression data. We estimate this quantity by creating a random labeling for the gene expression patterns in our data sets. As we can see from Figure 2 the distribution of scores in the randomized dataset is distinctly different than the distribution of scores for the original dataset. In particular, the errors scores achieved by the best scoring genes in the true data are extremely unlikely in random data.

Aside from the statistical significance of the selected genes, we would also like to evaluate their biological significance. To estimate this, we have ordered the genes in both data sets, according to their error score, and examined the genes at the top of the list (those that have low error score). Among the top 100 genes in the colon cancer data set there are a number of genes that are interesting from the perspective of a potential involvement in tumorigenesis including, for example, genes involved in cell cycle regulation and angiogenesis. There were also genes, for example (D63874) HMG-1 (human) and (T55840) tumor-associated antigen L6 (human), that have previously been found to have a particular association with colorectal carcinomas [23, 29].

Among the top scoring 137 clones in the ovarian cancer data, there are 85 clones that match to 8 genes that are cancer related (potential markers or expressed in cancer cells) and one that is related to increased metabolic rate (mitochondrial gene). These genes are keratin 18 (breast cancer), pyruvate kinase muscle 2 (hepatoma), thymopoietin (cell proliferation), HE4 (ovarian cancer), SLPI (many different cancers, among them lung, breast, oropharyngeal, bladder, endometrial, ovarian and colorectal carcinoma), ferritin H (ovarian cancer), collagen 1A1 (ovarian cancer, osteosarcoma, cervical carcinoma), and GAPDH (cancers of lung, cervix and prostate). In addition, 2 clones with no homology to a known gene are found in this selection. Given the high number of cancer related genes in the top 137, it is likely that these novel genes exhibit a

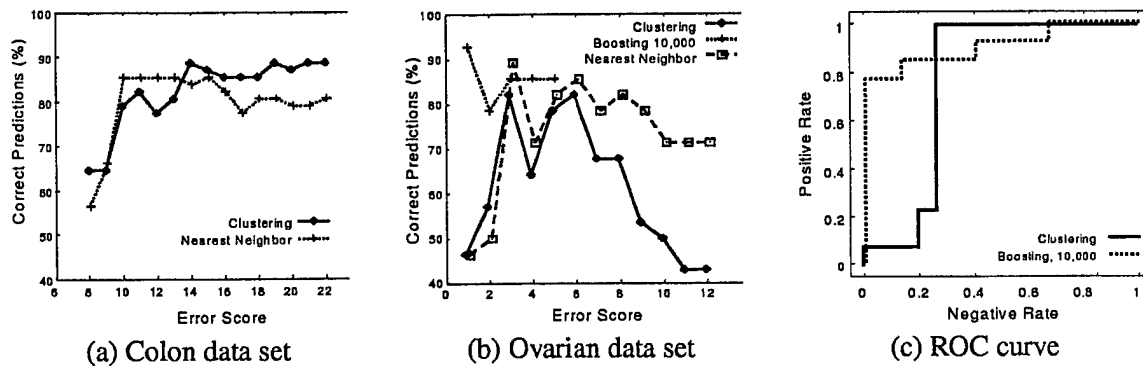


Figure 3: Curves showing how the classification accuracy depends on the threshold for selecting genes. The x -axis shows the error-score threshold for selecting genes. The y -axis shows classification accuracy based on LOOCV. Accuracy curves are (a) for colon data set, and (b) for ovarian data set. (c) shows ROC curves for two methods that are applied to the ovarian data set with error score threshold set to 3.

similar cancer-related behavior. We conducted expression validation for GAPDH, SLPI, HE4 and keratin 18 which confirmed the elevated expression in some ovarian carcinomas compared to normal ovarian tissues.

When using gene selection, we need to pre-process the training data to select genes. Then, the classification procedure is applied to the training data restricted to the subset of selected genes. The gene selection stage is given a parameter k , which determines the largest error-score allowed. It then selects all genes that have a smaller or equal error score on the training data.

To evaluate performance with gene selection, we have to be careful to evaluate together the two stage process of gene selection and classification. Thus, in each cross-validation trial, gene selection is applied based on the training examples in that trial. Note, that since the training examples are different in different cross validation trials, we expect the number of genes with error scores below a given threshold to change between trials.

Figures 3(a) and 3(b) show the classification accuracy for some of the methods we described above when we vary the maximal score distribution of genes we select. We note that SVM and boosting were also run with feature selection. However, due to technical issues, they were run on subsets of fixed sizes. The results for colon cancer show that both achieve approximately 80% accuracy with subsets of size 100 instances and bigger. Note also that for the clustering method this presentation is over pessimistic as it treat unclassified tissues as failures. For example, in the ovarian data set, for error score of 12, the clustering based classifier made 12 correct prediction, 5 wrong predictions, and 11 'unknown' predictions.

Both these graphs show that we can achieve quite a good classification performance with a small number of genes. For example, in the colon cancer data set, feature selection neither helps the clustering approach, nor does it significantly harm its behavior. We see that even for error threshold 10, which corresponds to selecting 10 genes on average, we see good prediction performance. In the ovarian data set, the critical threshold value is 3, which corresponds to selecting, on average, 173 clones.

In the colon data set, gene selection does not lead to significant improvement. On the other hand, in the ovarian data set, gene selection leads to impressive improvement in all methods. All three methods perform well in the region between threshold 3 (avg. 173 clones) to 6 (avg. 4375 clones). Note that Boosting performs well even with fewer clones. Figure 3(c) shows an ROC curve for Boosting and the Clustering approach with threshold of 3. As we can see, although both methods have roughly the same accuracy with this subset of genes, their ROC profile is strikingly different. These curves clearly show that the Clustering approach makes false positive errors, while the boosting approach makes false negative errors.

5 Sample Contamination

Cancer classification based on array-based gene expression profiling may be complicated by the fact that clinical samples, e.g. tumor vs. normal, will likely contain a mixture of different cell types. In addition, the genomic instability inherent in tumor samples may lead to a large degree of random fluctuations in gene expression patterns. Although both the biological and genetic variability in tumor samples have the potential

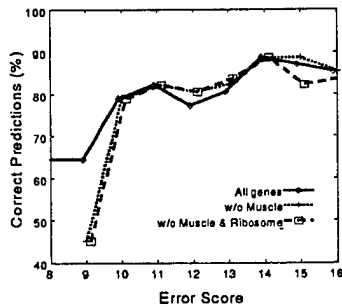


Figure 4: Curves showing the predictive performance of clustering methods in the original Alon et al. data set, and data sets where muscle specific, and ribosomal genes were removed. All estimates are based on LOOCV evaluation. These results show that even without the obvious contaminations, our methods are successful in reliably predicting tissue type.

to lead to confusing and difficult to interpret expression profiles, gene expression profiling does allow us to efficiently distinguish tumor and normal samples, as we have seen in the previous sections. However, the presence of different cell types within and between samples could lead to identification of genes that strongly affect cluster formation but which may have little to do with the process being studied, in this case tumorigenesis. For example, in the case of the colon cancer data set presented above, a large number of muscle-specific genes were identified as being characteristic of normal colon samples both in our clustering results and in the results of Alon et al. [1]. This is most likely due to a higher degree of smooth muscle contamination in the normal versus tumor samples.

This raises the concern that our classification may be biased by the presence of muscle specific genes. To test this hypothesis, we performed the following experiments. We listed the top 200 error-score ranking genes in the colon cancer data set, and identified muscle-specific genes. These include (J02854) myosin regulatory light chain 2, smooth muscle isoform (human); (T60155) actin, aortic smooth muscle (human); and (X12369) tropomyosin alpha chain, smooth muscle (human) that are designated as smooth muscle-specific by Alon et al.'s analysis, and (M63391) desmin (human), complete cds; (D31885) muscle-specific EST (human); and (X7429) alpha 7B integrin (human) which are suspected to be expressed in smooth muscle based on literature searches.

An additional form of "contamination" is due to the high metabolic rate of the tumors. This results in high expression values for ribosomal genes. Although such high expression levels can be indicative of tumors, such a finding does not necessarily provide novel biological insight into the process, nor provide a diagnostic tool since ribosomal activity is present in virtually all tissues. Thus, we also identified ribosomal genes in the top 200 scoring genes.

Figure 4 shows the performance of the clustering approach on three data sets: the full 2000 gene data set, a data set without muscle specific genes, and a data set without both muscle specific and ribosomal genes. As the learning curves show, the removal of genes affects the results only in cases using the smallest sets of genes. From error score threshold of 10 (avg. 9.1 genes) and higher, there is no significant change in performance for the procedure. Thus, although muscle specific genes can be highly indicative, the classification procedure performs well even without relying on these genes.

Although the muscle contamination did not necessarily alter the ability of this gene set to be used to classify tumor vs. normal samples in this case, it will continue to be important to account for possible affects of tissue contamination on clustering and classification results. Experimental designs that include gene expression profiles of tissue and/or cell culture samples representative of types of tissue contaminants known to be isolated along with different types of tumor samples (for example see Perou et al. [19]), can be utilized to help distinguish contaminant gene expression profiles from those actually associated with specific types of tumor cells.

6 Conclusions

In this paper we examined the question of tissue classification based on expression data. Our contribution is three-fold. First, we introduced a new cluster-based approach for classification. This approach builds on the recent development of clustering algorithms that are suitable for gene expression data. Second, we performed rigorous evaluation of this method, and a few known methods from the machine learning literature. These include large margin classification methods (SVM and stump-boosting) and the nearest-neighbor method. Third, we investigated the issue of gene selection in expression data. As our results for the ovarian data set show, a large number of clones can have a negative impact on predictive performance. We showed

that a fairly simple selection procedure can lead to significant improvements in prediction accuracy. In addition, we highlighted the issue of sample contamination and estimated the sensitivity of our approach to such contamination.

One clear future direction is extracting from the learned classifiers the genes that play a dominant role in them (i.e. those genes on which the classification relies the most). This might reveal some previously unknown disease related genes, which might point a direction for biological research.

References

- [1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA*, 96:6745–6750, 1999.
- [2] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 2000. to appear.
- [3] C. J. C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
- [4] S. Chu, J. DeRisi, M. Eisen, J. Mullholland, D. Botstein, P. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.
- [5] P. A. Clarke, M. George, D. Cunningham, I. Swift, and P. Workman. An analysis of tumor gene expression following chemotherapeutic treatment of patients with bowel cancer. In *Proc. Nature Genetics Microarray Meeting 99*, page 39, Scottsdale, Arizona, 1999.
- [6] C. Cortes and V. Vapnik. Support vector machines. *Machine Learning*, 20:273–297, 1995.
- [7] J. DeRisi, V. Iyer, and P. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 282:699–705, 1997.
- [8] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
- [9] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA*, 95:14863–14868, 1998.
- [10] Lander et al. to appear, 1999.
- [11] B. Everitt. *Cluster Analysis*. Edward Arnold, London, third edition, 1993.
- [12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sciences*, 55:119–139, 1997.
- [13] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J. Hudson, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P.O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.
- [14] Kim lab home page. <http://cmgm.stanford.edu/kimlab/>.
- [15] J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. M. Trent, and P. S. Meltzer. Gene expression profiling of *Alveolar rhabdomyosarcoma* with cDNA microarrays. *Cancer Research*, 1998.
- [16] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, pages 1137–1143. Morgan Kaufmann, San Francisco, Calif., 1995.

- [17] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Want, M. Kobayashi, H. Horton, and E. L. Brown. DNA expression monitoring by hybridization of high density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.
- [18] L. Mason, P. Bartlett, and J. Baxter. Direct optimization of margins improves generalization in combined classifiers. In *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, Mass., 1999.
- [19] C. M. Perou, S. S. Jeffrey, M. v de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. F. Lee, D. Lashkari, D. Shalon, P. O. Brown, and Botstein D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Nat. Acad. Sci. USA*, 96:9212–9217, 1999.
- [20] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.
- [21] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.
- [22] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 1999. to appear.
- [23] TH. Schiedeck, S. Christoph, M. Duchrow, and H.P. Bruch. Detection of h16-mrna: new possibilities in serologic tumor diagnosis of colorectal carcinomas. *Zentralbl Chir*, 123(2):159–162, 1998.
- [24] M. Schummer. in preparation, 1999.
- [25] M. Schummer, W. NG, Bumgarner R., Nelson P., B. Schummer, L. Hassell, L. Rae Baldwin, B. Karlan, and L. Hood. Comperative hybridization of an array of 21,500 overian cDNAs for the discovery of genes overexpressed in overian carcinomas. *Gene*, 1999. in print.
- [26] J. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.
- [27] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1999.
- [28] X. Wen, S. Fuhmann, G. S. Micheals, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Nat. Acad. Sci. USA*, 95:334–339, 1998.
- [29] Xiang YY, Wang DY, Tanaka M, Suzuki M, Kiyokawa E, Igarashi H, Naito Y, Shen Q, and Sugimura H. Expression of high-mobility group-1 mrna in human gastrointestinal adenocarcinoma and corresponding non-cancerous mucosa. *Int J. Cancer*, 74(1):1–6, Feb 1997.

A Support Vector Machines

A linear decision rule can be represented by a hyperplane in R^N such that all examples on the one side of the hyperplane are labeled positive and all the examples on the other side are labeled as negative. Such a rule can be represented by a vector $w \in R^N$ and a scalar b that together specify the hyperplane $w \cdot x + b = 0$. Classification for a new example x is performed by computing $\text{sign}(w \cdot x + b)$. Recall that $|\frac{x \cdot w + b}{\|w\|}|$ is the distance from x to the line $x \cdot w + b = 0$. Thus, if all points in the training data satisfy

$$l_i(x_i \cdot w + b) \geq 1 \tag{1}$$

then we know that they are all correctly classified, and all of them have a distance of at least $1/\|w\|$ from the hyperplane. We can find the hyperplane that maximizes the margin of error by solving the following quadratic program:

$$\begin{aligned} &\text{Minimize } \|w\|^2 \\ &\text{Subject to } l_i(x_i \cdot w + b) \geq 1 \text{ for } i = 1, \dots, m. \end{aligned}$$

Input:

- A data set of N labeled examples $\{(x_1, l_1), \dots, (x_N, l_N)\}$
- A weak learning algorithm L .

Initialize the distribution over the data set: $D_1(x_i) = 1/N$

For $t = 1, 2, \dots, T$

- Call L with distribution D_t ; Get back a hypothesis h_t .
- Calculate the error of h_t : $\epsilon_t = \sum_{i=1}^N D_t(x_i) 1(l_i \neq h_t(x_i))$
- Set $\alpha_t = \frac{1}{2} \log \frac{\epsilon_t}{1-\epsilon_t}$
- Set the new distribution to be:

$$D_{t+1}(x_i) = \frac{D_t \exp(-\alpha_t 1(l_i \neq h_t(x_i)))}{Z_t}$$

Where Z_t is a normalization factor, chosen so that D_{t+1} will sum to 1.

Output: The final hypothesis $h(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x))$

Figure 5: The AdaBoost algorithm.

Such quadratic programs can be solved in the dual form. This dual form can be posed in terms of auxiliary variables α_i . The solution has the property that

$$w = \sum_i \alpha_i l_i x_i,$$

and thus, we can classify a new example x by evaluating

$$\text{sign}\left(\sum_i \alpha_i l_i \langle x_i, x \rangle + b\right) \quad (2)$$

In practice there is a range of optimization methods that can be used for solving the dual optimization problem. See [3] for more details.

The SVM dual optimization problem and its solution have several attractive properties. First, only a subset of the training examples determine the position of the hyperplane. Intuitively, these are exactly those samples that are at the distance $1/\|w\|$ from the hyperplane. It turns out that the dual problem solution assigns $\alpha_i = 0$ to all examples that are not “supporting” the hyperplane. Thus, we only need to store the *support vectors* x_i for which $\alpha_i > 0$. (Hence the name of the technique.)

Second, the dual form of the quadratic optimization problem involves only cross-products of vectors in R^N . In other words, vectors x_i do not appear outside the scope of a cross-product operation. Similarly, the classification rule (2) only examines vectors in R^N inside the cross-product operation. Thus, if want to consider any projection $\Phi : R^N \mapsto R^M$, then we can find an optimal separating hyperplane in the projected space, by solving the quadratic problem with cross-products $\langle \Phi(x_i), \Phi(x_j) \rangle$.

In many cases, we can perform the optimization in high-dimensional spaces, by efficient computation of the cross-product in these spaces. A function $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is called a *kernel function*. For many projections, the kernel function can be computed in time that is linear in N , regardless of the dimension M .

B AdaBoost

AdaBoost algorithm was introduced in [12]. This algorithm is shown in Figure 5. See also [22] for a justification of the particular reweighting scheme used by Freund and Schapire.

Abstracts: NCI-EORTC Meeting, June 28 - July 1, 2000

HYBRIDISATION OF AN ARRAY OF 100,000 cDNAs WITH 32 TISSUES FINDS POTENTIAL OVARIAN CANCER MARKER GENES

SCHUMMER M¹, KIVIAT N¹, BEDNARSKI D¹, CRUMB GK¹, BEN-DOR A¹, DRESCHER C², HOOD L¹

¹University of Washington, Seattle; ²Swedish Medical Center, Seattle, USA

Ovarian cancer mortality could be largely reduced by early detection through a sensitive, specific and inexpensive serum assay. In order to find new markers, we used array technology to screen for genes with over-expression in the carcinomas vs. the normal tissues. An array of 102,680 clones, randomly selected from 3 unamplified ovarian cDNA libraries, was interrogated with probes from 32 well characterised tissues (normal ovaries, ovarian carcinomas, blood and liver). The hybridisation patterns were analysed with algorithms specifically created for such analysis. We found 2650 clones representing 883 genes with stronger expression in the tumours (476 matching known genes, 368 matching ESTs, and 48 novel genes). Some of the known genes were previously described cancer genes such as CD24, folate binding protein, c-myc, Her2/neu, mucin, metallothionein or c-jun. Detection of these genes demonstrates the power of our approach. To date, we performed real time-PCR-based expression validation on 34 novel and known genes in 72 tissues (18 normal ovaries, 40 ovarian tumors and cell lines), of which 20 genes were confirmed as over-expressed in the tumours. These genes are currently characterized by in-situ hybridisation on tissue sections, and by screening for antibodies and transcripts in patient sera that were collected with the tissues.

72 TISSUES (18 NORMAL OVARIES, 40 OVARIAN TUMORS AND CELL LINES), OF WHICH 20 GENES WERE CONFIRMED AS OVER-

Microarray-based gene profiling discovers potential ovarian cancer markers

Michèl Schummer 1*, Nancy Kiviat 2, Leroy Hood 1, Charles Drescher 3, Amir Ben-Dor 2,5, Zohar Yakhini 5, Martin McIntosh 3, Andrew F. Siegel 1,2, Irina Podolsky 1, Ingegerd Hellström 4, Karl-Erik Hellström 4, Nicole Urban 3

1 Institute for Systems Biology

2 University of Washington

3 Fred Hutchinson Cancer Research Center

4 Pacific Northwest Research Institute

5 Agilent Laboratories

* M.S., Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, WA 98105;
michel@drschummer.de

A.B-D. Agilent Laboratories, 3500 Deer Creek Rd, Palo Alto, CA 94304; amirbd@cs.washington.edu

Z.Y. Agilent Laboratories, 2 HaShlosa St, 67060 Tel Aviv, Israel; zohary@exch.hpl.hp.com

N.K. University of Washington, Box 359791, Seattle, WA 98195; nbk@u.washington.edu

N.U. Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, WA 98109;

nurban@fhcrc.org

M.M. Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, WA 98109;

mmcintos@fhcrc.org

C.D. Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, WA 98109;

cdresche@fhcrc.org

L.H. Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, WA 98105;

lhood@systemsbiology.org

I.P. Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, WA 98105;

ipodolsky@systemsbiology.org

I.H. Pacific Northwest Research Institute, 720 Broadway, Seattle, WA 98122; ihellstrom@pnri.org

K-E..H. Pacific Northwest Research Institute, 720 Broadway, Seattle, WA 98122; khellstrom@pnri.org

A.S. University of Washington, Box 353200, Seattle, WA 98195; asiegel@u.washington.edu

ABSTRACT

Ovarian cancer is the fourth most important cause of female cancer mortality resulting in over 14,000 deaths in the U.S. every year. Most women are diagnosed with the cancer in its late stage, resulting in survival rates of 25% and below. There are currently no markers for screening of the general population. Such markers could potentially increase survival through detection of the tumor when still confined to the ovary. Using an approach of comparing transcript abundance in tumors and normal tissues, notably multiple rounds of microarray interrogation and expression confirmation, we identified 12 genes with potential as markers for an ovarian cancer screening assay. We showed that the proteins WFDC2 (HE4) and MSLN (Mesothelin) are secreted into serum and can be detected in the blood of ovarian cancer patients. Further work is under way to establish whether WFDC2 and MSLN provide clinically useful markers, especially when tested for in combination with other serum assays in ovarian carcinoma patients.

INTRODUCTION

Ovarian cancer is the fourth most important cause of female cancer mortality resulting in over 14,000 deaths un the U.S. every year [American Cancer Society, 2000 #881]. Although 5 year survival is high for women diagnosed with ovarian cancer still confined to the ovary, 75% of women with ovarian cancer initially present with extra-ovarian disease [Ries, 1999 #882]. For this latter group, five year survival falls to below 30% [Ries, 1999 #882]. Development of sensitive, specific and inexpensive screening assays for use in the general population is therefore a priority for ovarian cancer control. Currently, screening is not recommended for use in the general population, but randomized controlled trials are now testing the efficacy of screening using transvaginal sonography (TVS) and the ovarian cancer marker CA125 [Kramer, 1993 #883; Jacobs, 1999 #876]. As a first line screen, each, CA125 and TVS, have relatively poor specificity [Jacobs, 1999 #876], which is improved by first using CA125 followed by TVS. However, because ovarian cancer is rare, specificity of 99.6% is needed to achieve a positive predictive value of 10% [Urban, 1999 #695]. Better markers are therefore needed. In a previous study, we identified WFDC2 (formerly HE4) as being a potential marker for ovarian cancer. We designed the present study as a collaborative effort between clinicians, molecular biologists, biostatisticians and computer scientists to expand and extend these results. Using multiple rounds of microarray interrogation and expression validation, we confirmed the correlation of WFDC2 expression with ovarian cancer and identified 11 additional genes as markers for ovarian cancer screening assays. We showed that WFDC2 and MSLN (mesothelin) proteins are secreted into serum and can be detected in the blood of ovarian cancer patients. Further work is under way to establish whether WFDC2 and MSLN provide clinically useful markers, especially when tested for in combination with other serum assays in ovarian carcinoma patients.

RESULTS

Generations of libraries and arrays: We first generated a colony-based membrane array containing 102,680 randomly selected clones from three ovarian cDNA libraries. Of the 102,680 clones placed on the membrane array, 97,803 grew, as assessed by hybridization with a probe targeting the vector sequence of the clones. These 97,803 clones corresponded to approximately 30,000 unique genes (crude extrapolation from sequencing of 300 clones, data not shown). These cDNA libraries were not normalized (i.e. multiple clones coding for the same housekeeping genes were present). Since our previous studies demonstrated such housekeeping messages are expressed at higher levels in cancer than non cancer tissues [Schummer, 1999 #633] and might obscure detection of overexpressed transcripts of interest, we next hybridized the arrays with a "housekeeping" probe (described below) to identify such clones, and removed these clones from further analysis. Using this approach 10,716 clones were eliminated, leaving 87,087.

Hybridization of arrays with cancers and controls. We next interrogated these arrays with ³³P-labeled first-strand cDNA probes from 32 tissues (13 malignant ovarian cancer tissues, 2 serous cystadenomas, 12 normal ovaries, liver, and 4 peripheral blood lymphocyte preparations). We determined the relative intensities using a combination of statistical methods such as t-tests (described below) and TNoM scoring [Hedenfalk, 2001 #709; Bittner, 2000 #878; Ben-Dor, 2000 #639; Ben-Dor, 2000 #877]. We found 84,436 clones to be expressed at lower levels in normal, as compared to tumor tissues. These clones were considered to be of no further interest (Table 1). We characterized the identity of the remaining 2,651 clones by 5' end sequencing which resulted in 2061 obtained sequences. We submitted these 2061 sequences for nr and esdb database searches and found 1542 as homologous to known genes, 366 matching only to sequences in the esdb database, and 45 to be novel sequences. In summary, the 1542 clones corresponded to 883 unique sequences, not accounting for alternative splicing or alternative termination which may increase this number.

Table 1 - Consecutive reduction of number of clones to be analyzed

total number of possible clones on the array	102,680
clones with colony growth	97,803
clones negative for "housekeeping" probe	87,087
clones identified by statistics as expressed higher in the tumors*	2,651
number of clones with readable sequence	2,061
of those: clones with homology to known genes	1,542
clones with homology to ESTs	336
clones with no homology	45

Determining a set of separating genes by supervised tissue clustering. To identify cancer-associated overexpressed genes among the 2651 clones identified in the previous step, we next performed leave-one-out cross validation (LOOCV) with supervised CAST clustering [Ben-Dor, 2000 #639; Ben-Dor, 1999 #685]. This method requires a training data set and an unknown sample. The dataset consisted of 29 tissues

(13 malignant tumor tissues and 16 normal ovarian and non-ovarian tissues). For the first round of analysis, the 28 tissues with either a "tumor" or a "normal" tag attached were clustered by the expression values of the 2651 clones using CAST with a threshold that optimizes the partition into tumors and normals. Using this calculated threshold, we introduced the expression data from the 29th tissue and recorded whether it classified as tumor or normal. This was repeated 28 times with a different tissue as the unknown. In each round, we recorded 500 clones with the lowest coefficient of variation (standard deviation divided by the mean) in the tumor cluster. We thus identified 39 clones which were common in all 29 experiments, including 23 known genes, 3 novel genes and 13 ESTs. The 23 known genes were CD24, COL1A1 (collagen type 1, alpha 1), COX7B (cytochrome c oxydase, type vii), FTH1 (ferritin H), GAPD, WFDC2 (HE4), KIAA0762, KRT18 (keratin 18), LDHA (lactate dehydrogenase), OVGP1 (oviductal glycoprotein), PKM2 (muscle pyruvate kinase), S100A11 (calgizzarin), S100A6 (calcyclin), SLPI, SSR4, TMPO (thymopoietin), 2 mitochondrial genes, 3 immunoglobulin genes, and 2 genes from the major histocompatibility complex. Among these 23 are 10 genes which have been reported to be associated with cancer: GAPD [Kim, 1998 #595; Schek, 1988 #602; Tokunaga, 1987 #604], S100A11 [Van Ginkel, 1998], S100A6 [Komatsu, 2000 #678; Van Ginkel, 1998 #693], CD24 [Yang, 1999 #661; Fogel, 1999]; WFDC2 (HE4) [Schummer, 1999 #633]; COL1A1 [Kauppila, 1996 #495]; FTH1 [Tripathi, 1996 #493]; SLPI [Garver, 1994 #614]; LDHA [Rutzky, 1982 #858], and KRT18 [Schaller, 1996 #859]. We considered these 23 genes our first harvest of potential ovarian cancer-related genes.

Identifying additional clones of interest by unsupervised clone clustering. We next performed unsupervised CAST cluster analysis [Ben-Dor, 1999 #685] to search for clones with expression patterns similar to the 10 cancer-related genes identified above. We found 38 additional genes of interest who formed clusters with these 10 genes, including the 29 known genes ACTB (beta actin), ADAM15 (MDC15), ELF3 (ESE-1), ENO (alpha enolase), ERBB2 (Her2/neu), FOLR1, GAB2, GPR39, IFI27, IGF2, IGFBP2, JUN, KRT8 (keratin 8), LCN2 (lipocalin 2), LY6E (RIG-E), MNAT1, MSLN (mesothelin), MUC1, MYC, PAX2, PLTP, SAS, SCYA2 (MCAF), SDC4 (syndecan 4), ST5, TACSTD2 (GA733-1), TLE4, TRC8, YWHAE (14.3.3 epsilon), 8 ESTs and 1 novel gene. Of the 29 known genes, 9 genes have been reported to be associated with cancer: ACTB [Naylor, 1992], ELF3 [Oettgen, 1997], FOLR1 [Toffoli, 1997], KRT8 [Martens, 1999 #688], LCN2 [Argani, 2001 #860], MSLN [Wang, 1999 #489; Scholler, 1999 #692], MUC1 [Dong, 1997 #445; Ho, 1993 #550], MYC [Csokay, 1993 #153], PAX2 [Davies, 1999 #690].

Glass array hybridization: Confirmation of overexpression of the 2,651 clones and identification of additional clones of interest. To confirm the overexpression of the 77 cancer-associated genes identified by membrane-based analysis, we constructed PCR product-based glass microarrays (this type of array offers higher signal-to-noise ratio). Our array contained 1536 clones, including 1067 clones corresponding to 883 genes identified above, 402 clones from our previous ovarian membrane arrays based studies [Schummer, 1999 #633], and 67 control genes. We hybridized this array with first stand Cy5-labeled cDNA probes from 64 tissues (31 malignant ovarian cancer tissues, 7 serous cystadenomas, 24 normal ovaries, liver, and 1 peripheral blood lymphocyte preparation) and a Cy3-labeled reference cDNA probe generated from a pool

of RNA from all 64 tissues. In contrast to the membrane hybridizations above where we generated absolute hybridization values, the glass array experiments resulted in ratios between the tissue and the reference probe. Based on these ratios, we ranked the clones with respect to their ability to discriminate normal from cancer using the 10th percentile of the area under the Receiver Operating Characteristic (ROC) curve (pAUC0.1, Pepe et al., Biometrics, accepted). The top 100 clones thus identified included 27 known genes, 22 ESTs and 7 novel sequences. Of the 27 known genes, 18 had been previously identified by our initial membrane-based approach (CD24, FOLR1, GAPD, GPR39, WFDC2, IFI27, KIAA0762, KRT18, KRT8, LCN2, MSLN, MUC1, OVGP1, S100A6, SDC4, SLPI, TACSTD2, YWHAZ). Likewise 10 of the 22 ESTs and 4 of the 7 novel sequences found on glass arrays had been previously identified by membrane-based analysis. In summary, using both the membrane and glass microarrays, we identified a total of 101 genes of interest as potential markers of ovarian cancer.

Further Confirmation of overexpression of the genes of interest in tissues by real-time PCR and in situ hybridization. We then used real-time PCR to further confirm the cancer-related overexpression of 59 genes (33 known genes, 23 ESTs and 3 novel genes) assayed on the arrays in a total of 128 tissues, not used on the previous array hybridization experiments. Those genes chosen for real-time PCR based confirmation included 1) all genes which were identified by both membrane and glass array hybridization (n=32), 2) all genes (identified by either membrane or glass arrays) which appeared to code for a secreted or membrane bound protein (n=23), and 3) all genes (identified by either) which had previously been reported to be cancer-associated (n=19). We thus confirmed the cancer-associated overexpression of WFDC2, MSLN, LCN2, ENO1, IFI27, KRT8 and, to a lesser degree, that of a novel gene (T000M-134-O23) and two ESTs (T000M-106-H01, T000M-31-N08)(Table 2). A tissue was regarded positive if its expression value was 3 standard deviations above the mean of all 73 normal tissues. For comparison, we assayed MUC16, the recently discovered gene [Yin, 2001 #879] coding for the ovarian cancer marker CA125. The genes WFDC2 and MSLN appeared to be among the markers with the best sensitivity (expression in ovarian carcinomas, especially of early stage) and specificity (no expression in blood cells and in normal ovaries, no or low expression in other normal tissues) for identification of ovarian cancers. We additionally confirmed their expression patterns in ascitic cells from patients with and without malignancy and with and without ovarian cancer and by in situ hybridization of 42 tissues (12 normal ovarian tissues, 8 benign ovarian tumors, 22 ovarian carcinomas). Lastly, WFDC2 and MSLN transcripts were both localized in the tumor cells and were absent from stromal cells, connective tissue cells XXX. (Nancy will add more about the in situ, including a picture)

Table 2 - Percentage of 128 tissues in 8 categories positive for potential marker genes

	ENO1	ELF3	FOLR1	IFI27	KRT8	LCN2	MSLN	SLPI	WFDC2	novel	EST	EST	MUC16
normal non-ovary, n=32	3%	6%	3%	0%	3%	3%	0%	3%	3%	0%	9%	0%	3%
PBL, n=7	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
normal ovary, n=34	6%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	2%	0%
benign tumor, n=11	9%	0%	9%	9%	0%	0%	27%	9%	0%	8%	8%	0%	27%
borderline, n=2	50%	50%	0%	0%	0%	0%	0%	50%	50%	0%	0%	0%	100%

late stage, n=31	23%	29%	32%	19%	3%	0%	39%	3%	19%	5%	0%	12%	65%
metastatic, n=6	17%	33%	50%	17%	0%	0%	33%	17%	33%	17%	0%	0%	33%

The table lists the number of tissues (in percent) in each tissue category (normal non-ovarian tissues, peripheral blood lymphocytes, benign ovarian tumors, borderline ovarian tumors, early, late stage and metastatic ovarian carcinomas) which were positive for each of the 12 potential marker genes and, for comparison, MUC16, the gene encoding CA125.

Confirmation of overexpression of proteins of interest in ascitic fluids and serum. Since our aim was to identify blood-based biomarkers for identification of women with ovarian cancer, we undertook a pilot study on a sample of 80 sera (40 from women with ovarian cancer and 40 from healthy controls) to determine whether the gene products of WFDC2 and/or MSLN could be detected in serum. Table 3 summarizes the sensitivities of these markers used in an enzyme-linked immunoabsorbent assay (ELISA). We used the maximum value observed in control sera to set thresholds, and found 18/40 (45%) cases having WFDC2 levels above its threshold and 26/40 (65%) having MSLN elevated above its threshold. When used together a total of 31=13+13+5 of 40 cases are elevated, for a combined sensitivity of 71%. Further assessment of these two markers on 445 sera is under way to establish whether WFDC2 and MSLN provide clinically useful markers, especially when tested in combination with other serum assays in patients with ovarian carcinoma. The 445 sera include 227 healthy controls, 53 ovarian cancer cases of various histologies, and 86 other cancers and benign conditions.

Table 3 - Results of the Sandwich ELISAs for WFDC2 and MSLN

		WFDC2		
		negative	positive	total
MSLN	negative	9	5	14
	positive	13	13	26
	total	22	18	40

Listed are the number of tissues negative or positive for WFDC2 (columns) and MSLN (rows), either for each marker separately or for both combined. A marker is positive if found elevated above the maximum concentration observed among the 40 control sera.

DISCUSSION

In the present study we used tissues from 57 ovarian cancers and 97 controls to examine 100,000 clones by cDNA array hybridization to identify genes that are expressed at high levels in ovarian cancer, but at low levels or not at all in normal tissues, including ovary. Using this approach we confirmed the importance of WFDC2 (HE4, found during our previous array-based studies) and identified an additional 59 genes of interest. The overexpression of these genes in ovarian cancer was then confirmed by both real-time PCR quantitation and in-situ hybridization, leaving 11 genes with higher transcript levels in the cancers. In addition, we demonstrated that the WFDC2 and MSLN (Mesothelin) protein could be detected in sera. Most importantly, our ELISA results on 40 ovarian cancer cases and 40 healthy controls suggest that both proteins have the potential to become clinically relevant ovarian cancer markers.

Other investigators have previously examined the expression patterns of multiple genes in ovarian cancers as compared to controls [Hough, 2000 #701; Martoglio, 2000 #866; Ismail, 2000 #865; Ono, 2000 #807; Welsh, 2001 #716]. However, the present study is unique with regards to both the large number of genes and benign and malignant ovarian and non-ovarian tissues examined. Further, this is the only study where the array-based elevated expression patterns were confirmed by real time PCR, in-situ hybridization or detection of protein in sera.

Ismail et al. identified 255 differentially expressed genes by examining ovarian cancer (CSOC) and human ovarian surface epithelial (HOSE) cell lines using representational differences analysis. They then used array analysis to define the expression patterns of these 255 genes in 5 additional HOSE and 10 additional CSOC cell lines. In general, it is difficult to compare the results of arrays when the arrays contain different genes, however, interestingly, 4 out of the 9 known genes found by these researchers as being overexpressed in the ovarian cancer cell lines (collagens, SPARC, ANXA3, OSF2) were also identified in our own study. Martoglio et al. probed an array containing 332 angiogenesis- and tumorigenesis-related genes with an mRNA-derived probe from 5 normal ovarian tissues and 4 poorly differentiated serous adenocarcinomas and identified several cancer associated overexpressed genes (VEGF, ANGPT1, MST1R, BAD, MYBL2, MSN and REST). None of these genes were identified as overexpressed in ovarian cancers by our study which likely reflects the fact that we examined many more genes, not just those associated with angiogenesis. It is therefore possible that these genes may not have been represented on our membrane arrays. Sequencing of a random sample of clones from each of the cDNA libraries from which we derived our arrays, did not reveal any of these genes to be present. In another study, Ono et al. examined the expression patterns of 9 ovarian adenocarcinomas as compared to tissues from the histologically normal contralateral ovaries. Using a microarray containing 9121 previously identified cancer-related genes these investigators identified 39 genes which were up-regulated in ovarian cancer. Four of these genes were also found in the present study (ADSS, WFDC2, KRT17 and KRT18). Welsh et al. examined the expression patterns of 49 ovarian tissues/cell lines using an oligonucleotide array (Affymetrix) containing 6000 human genes, and identified 30 cancer-associated overexpressed transcripts. Of these, 19 were also identified by

our study, including CD24, WFDC2, CD9, LU, TACSTD1, KRT18, KRT19, KRT8, UCP2 (homologue identified), PAX8, ENO1, ELF3, SNRPB (close family member identified), TPI1, COX6A1 (COX6B was identified), ASS, MUC1, KPNA2 and SPINT2. Lastly, Hough et al. used SAGE to compare transcript abundance in cDNA libraries generated from 3 ovarian cancers to cDNA libraries from an ovarian surface epithelium cell line (HOSE) and a benign cystadenoma cell line (ML10). They reported 23 known genes to be overrepresented in the cDNA libraries derived from the cancer tissues. Our present study found 5 of these 23 as overexpressed as well (WFDC2, MSLN, SLPI, MUC1, TACSTD2).

The strengths of the present study include the use of cDNA microarrays containing a large number of genes which were predominantly expressed in normal and malignant ovarian tissues. (Rephrase) This focus on genes expressed in ovarian tissues increased the likelihood of detection of genes which had not been previously identified as overexpressed in cancer. Another strength of this study, was the fact that membrane-based array findings were confirmed by glass arrays, with additional confirmation by real-time quantitative PCR, in situ hybridization and detection of protein. Shortcomings of the present study include our failure to detect MUC16, the gene for CA125 which is known to be present in the sera of approximately 75% of late stage ovarian cancers patients [Miller, 1994 #875]. Our failure to detect this gene may have been due to the absence of a cDNA clone containing MUC16 on the array. Random sequencing of clones from each library did not reveal the presence of a MUC16 clone. Furthermore, MUC16 was expressed in only 4 out of the 13 tumors used to interrogate the membrane arrays, as assayed by real-time PCR analysis.

We currently are generating monoclonal antibodies to other proteins found in this study. Furthermore, the performance of ELISA assays for WFDC2 and MSLN on a larger cohort of cases and controls will show whether these proteins hold up to their promising performance as markers for the detection of ovarian cancer.

MATERIALS AND METHODS

Specimen accrual. We obtained 202 tissue biopsies, ascites samples, or RNA from biopsies and matching sera according to the procedures approved by the institutional review boards of the University of Washington, Swedish Hospital and Fred Hutchinson Cancer Research Center, Seattle, WA. All malignant tissue (macrodissected free of stroma and necrosis) and serum samples were obtained from women at primary surgery. Histopathologic examination found tumor samples to be composed of more than 80% tumor cells. The RNA from ovarian surface epithelial cultures was obtained from B. Karlan, Cedars Sinai Hospital, Los Angeles, CA and R. Hernandez, University of Washington, Seattle, WA.

RNA preparation. Total RNA was prepared by TRIZOL (Invitrogen, Carlsbad, California) extraction followed by ion exchange column purification (RNeasy, Qiagen, Valencia, California) and LiCl precipitation. RNA levels were determined by photometry and the integrity of RNA confirmed on agarose gels.

Generation of representative cDNA array from ovarian tissues. From three cDNA libraries (normal ovary, n=3 specimens; primary late stage serous ovarian carcinoma, n=4; metastatic ovarian cancer, n=6), 102,680 clones were selected (9,216, 17,664 and 83,712 respectively) and arrayed in the form of colonies onto 32 sets of 5 nylon membranes. The colonies were lysed and the DNA fixed onto the membranes using a modified Southern blot protocol [Sambrook, 1989].

Membrane array hybridization. Membrane array hybridization was performed as described earlier [Schummer, 1999 #633]. The "housekeeping" probe consisted of the following genes: ACTB; COMT; EEF1A1; EEF1G; MTATP6; MTCO1-MTCO3; MTCYB; MTND1-MTND6; OVGP1; RPL3, 5, 6, 7, 7A, 9, 18, 27, 30; RPP0; RPS3, 3A, 4, 6, 11-14, 16-18, 21, 24, 25, 28; 18 S rRNA; 28 S rRNA. The identity of a subset of the 10,716 of clones which hybridized with the "housekeeping" probe was further confirmed by sequencing.

Generation of the glass array. The cDNA clones were amplified by PCR using primers recognizing the vector sequences outside the multiple cloning site. The PCR product was purified by size exclusion columns (Sephacryl S-500, AmershamPharmacia, Piscataway, New Jersey). The PCR products were spotted in 50% DMSO on Amersham Type 7 glass slides using a Molecular Dynamics Generation II microarray printer (AmershamPharmacia, Piscataway, New Jersey). Each array contained 3072 positions with 1536 duplicated clones comprising 1469 PCR products from ovarian cancer libraries and 67 control genes (Pseudomonas and Arabidopsis clones, GFP, polyA DNA, empty positions and vector sequences).

Glass array hybridization. For each tissue, 100 µg of total RNA were reverse transcribed into first-strand cDNA as described earlier [Geiss, 2000 #681]. We generated a reference probe from a pool of equal amounts of total RNAs from all 64 tissues used for hybridization.

Statistical analysis. We used one-sided small-sample two-sample (16 normal tissues and the 23 tumors) t tests (one test for each of the 97,803 clones), using a critical t value of 4. We used z tests for the difference between the average expression level of 16 normal tissues and the single clone expression level for that tumor tissue, resulting in 2,249,469 tests, computed as 97,803 clones times 23 tumors, using a critical value of 10.09. We selected all clones with $t > 4$, or with $z > 10.09$ in at least one tumor, or with mean tumor expression > 2.5 times mean normal ovarian expression, or with mean tumor expression > 2.5 times mean normal expression (including ovarian and non-ovarian), or with average z score > 1.4 .

Clustering, classification and relevance scoring. The CAST algorithm [Ben-Dor, 2000 #639; Ben-Dor, 2000 #877] was performed in both supervised and unsupervised mode. For classification, we also used a CAST-based procedure as briefly discussed above and as explained in [Ben-Dor, 2000 #639]. TNoM scoring [Ben-Dor, 2000 #877] was used to identify tumor overexpressed clones, in addition to the t-test approach described above.

pAUC(0.1) The pAUC(0.1) statistical analysis was performed as described in Pepe et al (Pepe, Garnet and Schummer, Biometrics, accepted).

Real-time quantitative PCR validation. Total RNA was reverse transcribed using oligo-dT primer and Superscript II Reverse Transcriptase (Invitrogen, Carlsbad, California). Real-time quantitative PCR was performed on 128 tissues in duplicate using an ABI7700 machine (Applied Biosystems, Foster City, California) and the SYBR-Green protocol.

ELISA. The MSLN ELISA assay was designed earlier [Scholler, 1999 #692]. For WFDC2, we generated fusion proteins with either an immunoglobulin tail. Monoclonal antibodies were obtained to two different epitopes of the WFDC2 antigen. They were used to construct a double determinant ("sandwich") ELISA, analogous to a previously described one [Scholler, 1999 #692]. Sandwich ELISA assays were performed on sera from 40 ovarian cancer cases and 40 healthy controls. In both assays, the maximum cutoff of the 40 control sera was used to determine the threshold for each marker.

GenBank accession numbers

The novel sequence found as overexpressed in ovarian carcinomas, BI740074; EST sequences, AW166855, AI983043; TMP21, U61734.

ACKNOWLEDGEMENTS

Andrew F. Siegel holds the Grant I. Butterbaugh professorship at the University of Washington. We thank Roger Bumgarner for his microarray spotfinding software.

Appendix E
Project Two: Figures and Tables

Fig 1. Example of an immunoreactive phage plaque from a primary SEREX screen. The arrowhead indicates a single immunoreactive plaque (dark halo) amongst several hundred non-reactive plaques (clear spots).



Fig 2. Secondary SEREX screening of ovarian cancer cDNA clones by phage array. The left and right panels show duplicate nitrocellulose membranes containing a 2-D array of recombinant phage clones that were identified in a primary SEREX screen of an ovarian tumor cDNA library. The left panel was immunoblotted with serum from an ovarian cancer patient (stage III, serous) whereas the right panel was immunoblotted with serum from a normal control. Membranes were then probed with a human IgG-specific, AP-conjugated secondary antibody and developed with NBT/BCIP. Immunoreactive phage plaques appear as dark circles, whereas non-reactive phage are clear. The arrows indicate 6 phage that showed a cancer-specific pattern of immunoreactivity with these and other serum samples.

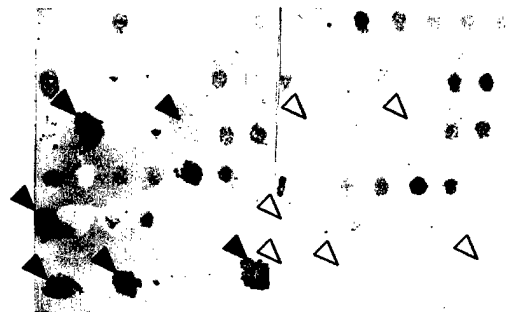


Fig 3. Western blot showing expressing of His-tagged recombinant tumor antigens in mammalian COS7 cells. Cells were transiently transfected with pcDNA3-based expression vectors encoding His-tagged ESO-1, p53 or Lac Z (as a control). Mock transfected cells served as a negative control. Nuclear extracts were prepared, subjected to SDS-PAGE and immunoblotted with a monoclonal antibody to the His tag (Sigma). Antibody detection was by enhanced chemiluminescence. Recombinant proteins are indicated by open arrowheads.

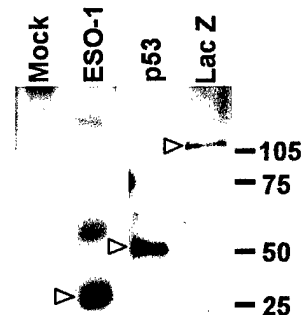


Fig 4. ELISA demonstrating serum antibody responses to p53 and ESO-1 in patients with ovarian cancer. Lysates from COS7 cells (see Fig. 3) expressing His-tagged p53, ESO-1 or, as a negative control, Lac Z were added to nickel-coated ELISA plates. After unbound proteins were washed away, serum from 10 ovarian cancer patients was added at 1:50 dilution, followed by HRP-conjugated goat anti-human IgG secondary antibody. Plates were developed with TMB and read at 450 nm. Patients #1-3 show a serum antibody response to p53, whereas patients #4-7 show a response to ESO-1. Patients #8-10 show no response to either protein.

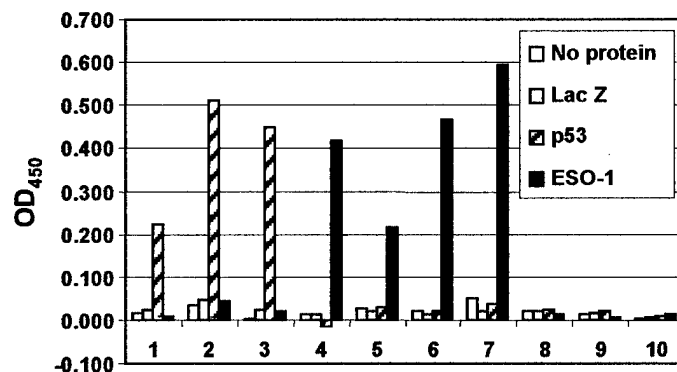


Table 2. List of antigens discovered by SEREX immunoscreening, and frequency of antibody responses to these antigens among patients with ovarian cancer and normal controls.

Antigen	Ovarian cases (n=54)	Controls (n=20)
p53	6	0
TOP2a	3	0
RUVBL	2	0
KIAA0035	1	0
HCAP-G	1	0
DLD	1	0
DDX9	1	0
STMN1	1	0
ILF3	1	0
NY-ESO-1	10	0
UBQLN1	3	0
HOXB6	3	0
ZFP161	1	0
HIS1	2	0
SPARC	1	0
CD44	1	0
YB-1	1	0
FBXO21	1	0
FLJ20267	2	0
DDX5	2	0
FLJ22318	1	0
KNSL6	2	0
FLJ10534	1	0
NKTR	1	0
IFI27	2	0
HSP40	1	0

Serum samples

Antigens	Serum samples																												Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
NY-ESO-1																													5
HOX B6																													2
HIS 1																													1
UBQLN1																													1
p53																													4
TOP2a																													1
KNSL6																													1
FLJ20627																													1
Total:	0	1	1	0	0	2	1	1	0	3	0	0	0	0	0	0	3	0	1	0	0	0	1	0	0	0	0	2	

Table 3. Pattern of IgG serum antibody responses to a subset of SEREX-defined antigens among 28 late-stage serous ovarian cancer patients. Each column shows the results for one ovarian cancer patient (numbered 1-28). Black cells indicate a positive response, as detected by SEREX-based arrays. All antigens were negative when tested against a panel of 20 normal control sera from age-matched women.

Appendix F
Project 2: Related Publications

- “MAGE-F1, a novel ubiquitously expressed member of the MAGE superfamily”

MAGE-F1, a novel ubiquitously expressed member of the *MAGE* superfamily

Brad Stone^{a,*}, Michel Schummer^b, Pamela J. Paley^c, Meghan Crawford^a, Molly Ford^a,
Nicole Urban^d, Brad H. Nelson^{a,e}

^aVirginia Mason Research Center, 1201 Ninth Avenue, Seattle, WA 98101-2795, USA

^bDepartment of Molecular Biotechnology, P.O. Box 357730, University of Washington, Seattle, WA, 98195, USA

^cDepartment of Obstetrics and Gynecology, P.O. Box 356460, University of Washington, Seattle WA, 98195-6406, USA

^dCancer Prevention Research Program, Fred Hutchinson Cancer Research Center MP-900, 1100 Fairview Ave. N.e Seattle, WA 98104, USA

^eMarsha Rivkin Center Department of Immunology, University of Washington, Seattle, WA 98195, USA

Received 30 October 2000; received in revised form 8 February 2001; accepted 19 February 2001

Received by T. Sekiya

Abstract

Most known members of the *MAGE* superfamily are expressed in tumors, testis and fetal tissues, which has been described as a cancer/testis or 'CT' expression pattern. We have identified a novel member of this superfamily, *MAGE-F1*, which is expressed in all adult and fetal tissues tested. In addition to normal tissues, *MAGE-F1* is expressed in many tumor types including ovarian, breast, cervical, melanoma and leukemia. *MAGE-F1* is encoded on chromosome 3, identifying a sixth chromosomal location for a *MAGE* superfamily gene. The coding region of *MAGE-F1* is contained within a single exon and includes a microsatellite repeat. Sequence analysis and expression profiles define a new class of ubiquitously expressed *MAGE* superfamily genes that includes *MAGE-F1*, *MAGE-D1*, *MAGE-D2/JCL-1* and *NDN*. The finding that several *MAGE* genes are ubiquitously expressed suggests a role for *MAGE* encoded proteins in normal cell physiology. Furthermore, potential cross-reactivity to these ubiquitously expressed *MAGE* gene products should be considered in the design of *MAGE*-targeted immunotherapies for cancer. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Tumor; *MAGE*; Ovary

1. Introduction

1.1. *MAGE* family genes

The gene encoding *MAGE-A1* was originally identified as a target antigen of patient-derived CD8⁺ T cells in melanoma (van der Bruggen et al., 1991). The identification and characterization of *MAGE-A1* was significant for two reasons. First, *MAGE-A1* exhibited a cancer/testis (CT) expression pattern, being detectable only in testis, fetal tissues and a subset of human tumors. Since the testis is an immune privileged site, and the germ cells expressing CT antigens are class I negative, this expression pattern is, from an immuno-

logical perspective, tumor-specific. Second, these efforts firmly established that cancer patients do in fact mount immune responses to tumor-specific gene products, or tumor antigens. These findings re-invigorated efforts to develop immunotherapies targeting *MAGE-A1* and other tumor-specific antigens for the treatment of human cancer.

Since the sequence of *MAGE-A1* was reported, 24 additional *MAGE*-related genes have been identified (Lurquin et al., 1997; Lucas et al., 1998, 2000; De Plaen et al., 1994; Pold et al., 1999; Gure et al., 2000). While some of these additional *MAGE* genes appear to be pseudogenes, others encode bona fide CT antigens and are immunogenic in human cancer (van der Bruggen et al., 1991; Van den Eynde and Boon, 1997; van Baren et al., 1999). Expression of *MAGE* proteins of the CT class is generally considered an inadvertent consequence of altered DNA methylation patterns in tumor cells, and in normal testis (De Smet et al., 1999; Serrano et al., 1996). The expression of CT class *MAGE* proteins by MHC Class I- and Class II-positive tumors appears to present the immune system with neo-auto-antigens to which tolerance has not

Abbreviations: CT, cancer/testis; MTE, multiple tissue expression; ORF, open reading frame; PBL, peripheral blood lymphocytes; RT-PCR, Real-time reverse-transcriptase PCR; STS, sequence tagged sites; *T_A*, annealing temperatures; UTR, untranslated region

* Corresponding author. Tel.: +1-206-223-6907, ext. 6-2714; fax: +1-206-223-7543.

E-mail address: bstone@vmresearch.org (B. Stone).

been established. Hence, a subset of cancer patients have been found to mount T and/or B cell responses to *MAGE* proteins (van der Bruggen et al., 1991; Van den Eynde and Boon, 1997; Tureci et al., 1997, 1999; Huang et al., 1999; Zorn et al., 1999). One unresolved issue is how tumor cells expressing immunogenic *MAGE* proteins escape elimination in immunocompetent patients.

Recently, two groups reported the identification of *MAGE* genes, *MAGE-D1* and *MAGE-D2/JCL-1*, that are expressed ubiquitously (De Plaen et al., 1994; Pold et al., 1999). In addition, several groups have shown that the distant *MAGE* relative *NDN* is expressed in post-mitotic neurons and many other normal tissues (Jay et al., 1997). These genes have not been shown to be aberrantly expressed or immunogenic in human cancer. In this report, we identify a fourth ubiquitously expressed member of the *MAGE* superfamily, *MAGE-F1*, that is expressed in normal somatic and tumor cells and is encoded at a novel location on chromosome 3. We also propose that an additional *MAGE* family gene, *KIAA1114*, is likely to be ubiquitously expressed. Collectively, these data suggest that at least five members of the *MAGE* superfamily are expressed in normal adult tissues, thereby defining a new class of *MAGE* genes. The expression of multiple *MAGE* family genes in normal somatic cells suggests members of this family play a functional role in normal cellular physiology.

2. Methods

2.1. Construction of an ovarian tumor cDNA library

RNA isolated from ten stage III/IV serous ovarian tumors was pooled and poly-A selected using an mRNA Separator kit from Clontech. Selected mRNA was converted to cDNA with a modified ZAP cDNA synthesis kit (Stratagene) and cloned into lambda TriplEx (Clontech).

2.2. Sequencing and analysis

Sequencing templates were prepared using QIAprep mini spin columns according to the manufacturers instructions. Sequencing was carried out using ABI BigDye sequencing reagents. Both ends of clone 38.1.1 (encoding *MAGE-F1*) were sequenced using the following vector primers; TCCGAGATCTGGACGAGC (sense primer) and TAATACGACTCACTATAGGG (anti-sense primer). Sequences were analyzed using BLASTn searches against NCBI (<http://www.ncbi.nlm.nih.gov/>) nr, EST and Unigene databases. All ESTs in Hs.84661 were downloaded and a provisional consensus sequence assembled using the Editseq and Seqman programs in the Lasergene software suite. The provisional consensus sequence was used in a blastn search of the NCBI human EST database to identify the following IMAGE clones: IMAGE2047755, IMAGEAW161349, IMAGE2422647, IMAGE35398 and IMAGE878566. These clones were purchased from GenomeSystems,

completely sequenced and assembled using the tools and protocols described above.

2.3. Rodent/human chromosome specific PCR

DNAs from cell hybrids hosting individual human chromosomes X, 14 and 3 were purchased from Coriell cell repositories (Mapping Panel #2). Primers for the following sequence tagged sites (STS) were synthesized and used as published in GenBank: DXS995, D14S1066, D3S1580, D3S1553 and WI-13826. In addition, primers amplifying overlapping fragments of the complete *MAGE-F1* coding sequence were synthesized and used with the indicated annealing temperatures (T_A). Primers amplifying *MAGE-F1* base 500 through 1210 were GGAGACTTGAA-GATTCTG (sense) and AGTCACACGATTTCTAC (anti-sense), $T_A = 53^\circ\text{C}$. Primers amplifying *MAGE-F1* bases 177 through base 610 were ATGTTGGAGACACCAGAGAG-CAGG (sense) and CTCCAGAGGTTTTAGTTTGTT-GATC (antisense), $T_A = 62^\circ\text{C}$. These two primer sets were used with the following amplification profile: 35 cycles at 94°C for 30 s, annealing for 30 s (at the indicated temperatures) and 72°C for 1 min. The fragment from bases 177 through 610 required the addition of 5% DMSO. Amplification products were resolved on 2% agarose gels and visualized with ethidium bromide staining.

2.4. Multiple tissue expression (MTE) and Northern blots

The MTE membrane was purchased from Clontech. RNA for Northern blotting was isolated from one normal human ovary and seven ovarian tumor specimens that had been flash frozen in liquid nitrogen immediately after surgical excision. Frozen tissues were crushed with a steel mortar and pestle, solubilized and extracted with RNA STAT-60 (TEL-TEST Inc.) according to the manufacturers instructions. PBL were solubilized and extracted without freezing. Ten micrograms of total RNA from each tissue was resolved on a formaldehyde/1% agarose gel, transferred overnight to Nylon membranes by capillary action and UV crosslinked (Manniat's). Both MTE and Northern membranes were probed using the protocol provided with the MTE blotting kit. Probes were synthesized from 50 ng gel purified PCR products representing the entire *MAGE-F1* coding region and the *Ubiquitin* control probe supplied with the MTE membrane. The key for the Clontech MTE blot is given in Table 1 and is also available online from Clontech.

2.5. Real-time PCR

RNA was extracted from cells or tissues by the TRIZOL method followed by Qiagen RNeasy column purification. RNA quality was assessed on non-denaturing 1% agarose gels. Ten micrograms of RNA was reverse transcribed using Superscript (Life Technology) and 20 μM oligo(dT)19V according to the manufacturer's instructions.

Table 1
Key for the Clontech MTE blot

1	2	3	4
A Whole brain	Cerebellum, left	Substantia nigra	Heart
B Cerebral cortex	Cerebellum, right	Accumbens nucleus	Aorta
C Frontal lobe	Corpus callosum	Thalamus	Atrium, left
D Parietal lobe	Amygdala	Pituitary gland	Atrium, right
E Occipital lobe	Caudate nucleus	Spinal cord	Ventricle, left
F Temporal lobe	Hippo-campus		Ventricle, right
G p.g. Of cerebral cortex	Medulla oblongata		Inter-ventricular septum
H Pons	Putamen		Apex of the heart
5	6	7	8
A Esophagus	Colon, transverse	Kidney	Lung
B Stomach	Colon, descending	Skeletal muscle	Placenta
C Duodenum	Rectum	Spleen	Bladder
D Jejunum		Thymus	Uterus
E Ileum		Peripheral blood leukocyte	Prostate
F Ileocecum		Lymph node	Testis
G Appendix		Bone marrow	Ovary
H Colon ascending		Trachea	
9	10	11	12
A Liver	Leukemia, HL-60	Fetal brain	Years; total tma
B Pancreas	Hela S3	Fetal heart	Yeast tma
C Adrenal gland	Leukemia, K-562	Fetal kidney	<i>E. Coli</i> rna
D Thyroid gland	Leukemia, MOLT-4	Fetal liver	<i>E. Coli</i> DNA
E Salivary gland	Burkitt's lymphoma, Raji	Fetal spleen	Poly r(A)
F Mammary gland	Burkitt's lymphoma, Daudi	Fetal thymus	Human Cot-1DNA
G	Colorectal adeno-carcinoma, SW480	Fetal lung	Human DNA 100 ng
H	Lung carcinoma A549		Human DNA 500 ng

Following reverse transcription, reactions were diluted to 500 μ l with water. Each 20 μ l PCR reaction received 2 μ l of diluted cDNA with 0.2 units of Bioline Taq polymerase (Bioline), 3.0 mM $MgCl_2$, 2.4 mM dNTP, 1.34×10^{-4} X concentration SYBR Green (Molecular Probes), the supplied reaction buffer and 2.4 μ M of each primer. The reaction profile was 94°C for 60 min; 94°C for 25 min, 60°C for 25 min, 72°C for 45 s (40 X cycles). Amplifications were performed in a 96-well format with internal standards

included on each plate. Standard reactions included duplicate amplifications of the housekeeping gene *S31iii125* (GenBank accession number: U61734, primers listed below) from serial twofold dilutions of white blood cell cDNA. Non-reverse-transcribed RNA from an ovarian tumor was used as a negative control for all primer sets. All amplifications were performed in duplicate. Since SYBR green detects all double-stranded DNA, including potential amplification artifacts, the amplification products from each well were analyzed on an agarose gel for the presence of a single band of the appropriate size. The results for each 96-well plate were analyzed using the Sequence Detector program. This program computes expression levels relative to the standard *S31iii125* amplifications from the cDNA dilution series included on each plate. Each cDNA sample was analyzed with the following primers specific for beta-actin or *MAGE-F1*.

MAGE-F1:

- forward GGTTCGTGGCCAACTGCATA
- reverse CCCCTGGAACCAGATCATCAT

Beta-actin:

- forward ACTTCGAGCAAGAGATGGCCAC
- reverse CCTGTGTGGACTTGGGAGAGGA

S31iii125:

- forward CGACGCTTCTTCAAGGCCAA
- reverse ATGGAAGCCCAAGCTGCTGA

3. Results and discussion

3.1. Identification of *MAGE-F1*, in an ovarian tumor cDNA library

While screening an ovarian tumor cDNA expression library by SEREX (Sahin et al., 1995) with serum from ovarian cancer patients, we identified a cDNA clone with homology to the *MAGE* superfamily of tumor antigens. While this clone did not ultimately prove to be immunogenic with the sera used for screening, we noted it encoded a cDNA that was identical to several *MAGE*-like cDNA sequences grouped within a single unigene cluster, Hs.84461. We combined the sequence of the SEREX-derived cDNA with the ESTs clustered in Hs.84461 to assemble a preliminary consensus sequence. This sequence was used to identify five IMAGE consortium clones containing large inserts. Complete sequencing of these clones revealed a 1636 bp mRNA (Fig. 1). The largest open reading frame (ORF) was 927 bp and was flanked by a 176 bp 5' untranslated region (UTR), a 533 bp 3'UTR, and a poly-adenylation signal at position 1614.

I GCGGSCGCAGGTTTACTGCTCCGTTGCGGTGCGGCCAGCAGCCACAAAGCTCCC 53
 GCTGCCATTGCTCCYTGTA TCTCCCGCGTCACTGCCGCTGTCCAACCCCTCCCCGGGGCTTGCGCGCGCGGCTC 127
 CCACACCCCTCGGGCCGTGTACGCGCTCTGCACCTGCCTGCCGAAAAAC ATG TTG CAG ACA CCA GAG AGC 197
 M L Q T P E S 7
 AGG GGG CTC CCG GTC CCG CAG GCC GAG GGG GAG AAG GAT GGC GGC CAT GAT GGT GAG ACC CGG GCC 263
 R G L P V P Q A E G E K D G G H D G E T R A 29
 CCG ACC GCC TCG CAG GAG CGC CCC AAG GAG GAG CTT GGC GCC GGG AGG GAG GAG GGG GCT GCG GAG 329
 P T A S Q E R P K E E L G A G R E E G A A E 51
 CCC GCC CTC ACC CGG AAA GGC GCG AGG GCC TTG GCG GCC AAA KCC TTG GCA AGG CGC AGG GCC TAC 395
 P A L T R K G A R A L A A K A/S L A R R R A Y 73
 CGC CGG CTG AAT CGG ACG GTG GCG GAG TTG GTG CAG TTC CTC CTG GTG AAA GAC AAG AAG AAG AGT 461
 R R L N R T V A E L V Q F L L V K D K K K S 95
 CCC ATC ACA CGC TCG GAG ATG GTG AAA TAC GTT ATT GGA GAC TTG AAG ATT CTG TTC CCG GAC ATC 527
 P I T R S E M V K Y V I G D L K I L F P D I 117
 ATC GCA AGG GCC GCA GAG CAT CTG CGG TAT GTC TTT GGT TTT GAG CTG AAA CAG TTT GAC CGC AAG 593
 I A R A A E H L R Y V F G F E L K Q F D R K 139
 CAC CAC ACT TAC ATC CTG ATC AAC AAA CTA AAA CCT CTG GAG GAG GAG GAG GAG GAG GAG GAG CTG 659
 H H T Y I L I N K L K P L E E E E E E E E D L 161
 GGA GGA GAT GGC CCC AGA TTG GGT CTG TTA ATG ATG ATC CTG GGC CTT ATC TAT ATG AGA GGT AAT 725
 G G D G P R L G L L M M I L G L I Y M R G N 183
 AGC GCC AGG GAG GCC CAG GTC TGG GAG ATG CTG CGT CGG TTG GGG GTG CAA CCC TCA AAG TAT CAT 791
 S A R E A Q V W E M L R R L G V Q P S K Y H 205
 TTC CTC TTT GGG TAT CCG AAG AGG CTT ATT ATG GAA GAT TTT GTG CAG CAG CGA TAT CTC AGT TAC 857
 F L F G Y P K R L I M E D F V Q Q R Y L S Y 227
 AGG CGG GTG CCT CAC ACC AAT CCA CCA KCA TAT GAA TTC TCT TGG GGT CCC CGA AGC AAC CTG GAA 923
 R R V P H T N P P A/E Y E F S W G P R S N L E 249
 ATC AGC AAG ATG GAA GTC CTG GGG TTC GTG GCC AAA CTG CAT AAG AAG GAA CCG CAG CAC TGG CCA 989
 I S K M E V L G F V A K L H K K E P Q H W P 271
 GTG CAG TAC CGT GAG GCC CTA GCA GAC GAG GCC GAC AGG GCC AGA GCC AAG GCC AGA GCT GAA GCC 1055
 V Q Y R E A L A D E A D R A R A K A R A E A 293
 AGT ATG AGG GCC AGG GCC AGT GCT AGG GCC GGC ATC CAC CTC TGG TGA GGGTTGGTGAAAAGTTGGCC 1123
 S M R A R A S A R A G I H L W *** 308
 AGTGGGTCCCCGTGAGGACGAAC TACTGTCTGAGTCATAAGTAATATGGGTGGGGCGAGGGTCTTATTTCTGTA 1198
 GAAATCGTGTGACTTTAAGGATTAGATTTTGTATCTTATGTTTTGTAACATTTAATAATTACTGTTAAAAATGCTGT 1275
 TTGTAATAGAGATTGGTCTACTTTTCTGTAGGATTTTATTGTAGAGTTTGTCTGGTTTGTAAAAATGGATGGAA 1351
 GAAC TTGTATTATACTGTGATTTTGAACAGATTATGCAACATTGGAAGGAAGGCTGTACTTTGATGGTTTGAA 1426
 GGAAC T CAGCAGTATGATGATCTGGTTCCAGGGGAAAAAAATAGCTGGTTGGTGTCTAGCCCCCAACACTTTT 1500
 GTCTGTGTGTATAAAAGAAAGAAAGACTGGCATGTACCTTCATTGCTTAGCTATTGAGTATCTAGAGAAAAA 1574
 TTAATAATGCAATGAGTTAGCAGTATACCCTGGCACACTTAATAAATTAAACATTGTGGAGC 1636

The 927 bp ORF encodes a protein of 308 amino acids.

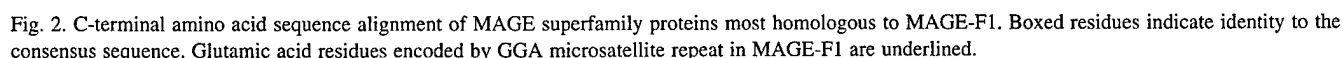


Table 2
Chromosomal location and tissue expression of MAGE superfamily genes

Gene	Chromosome	No. of amino acids	No. of EST entries ^a	Expression pattern	Identity to C-Terminus of MAGE-F1 ^b (%)
MAGE-A1	Xq28	308	1	CT	33
A2	Xq28	312	4	CT	35
A3	Xq28	314	9	CT	32
A4	Xq28	317	10	CT	34
A5	Xq28	124	1	CT	8
A6	Xq28	314	6	CT	33
A7	Xq28	pseudogene	0	NA	NA
A8	Xq28	234	3	CT	21
A9	Xq28	315	3	CT	33
A10	Xq28	369	12	CT	35
A11	Xq28	319	4	CT	33
A12	Xq28	314	0	CT	35
B1	Xp21	347	1	CT	38
B2	Xp21	319	5	CT	40
B3	Xp21	346	2	CT	39
B4	Xp21	346	0	CT	40
B5	Xp21	275	0	CT	34
B6	Xp21	407	0	CT	37
C1	Xq26	936	1	CT	36
C2	Xq26	373	0	CT	36
C3	Xq26	346	0	CT	32
D1	Xp11	573	238	Ubiquitous	45
D2/JCL-1	Xp11	606	283	Ubiquitous ^c	51
KIAA1114	Xp11 ^d	1383	69	Unknown ^c	49
NDN	15	321	59	Post mitotic Neurons ^c	41
E1	3	308	77	Ubiquitous	NA

^a The no. of EST entries is as of 4/1/00.

^b Amino acid identity to MAGE-F1 protein from residue 69 through 308 determined following optimal alignment with Lasergene Megalign Clustal program. No gap penalty is included in this score.

^c Broad tissue distribution in EST database, including libraries constructed from normal tissues other than testis.

^d KIAA1114 is an undescribed gene with multiple splice variants. The assignment of KIAA1114 to Xp11 is provisional, (see note in Section 3.3, paragraph 2).

BLASTp analysis revealed extensive homology with the MAGE superfamily of tumor antigens. The MAGE proteins with the greatest homology to the novel MAGE protein reported here were MAGE-D2/JCL-1, the human homologue of the murine necdin gene NDN, KIAA1114 and MAGE-D1 (Fig. 2 and Table 2). Homologies to these proteins (measured by amino acid identity) range from 51% (D2/JCL-1) to 41% (D1) within a 226 residue region in the C-terminal region of MAGE-F1. Homologies with MAGE superfamily proteins from the A, B and C subfamilies range from 22% (MAGE-A8) to 40% (MAGE-B4) within this same region. Analysis of the protein sequence with the BLOCKS algorithm revealed only the MAGE family motif (Henikoff and Henikoff, 1994). Based on these homologies and the novel chromosomal location described below, we propose the name *MAGE-F1* for this gene.

One unique feature of the *MAGE-F1* coding sequence is the presence of a trinucleotide microsatellite repeat, GGA, beginning at position 643 (Fig. 1). This repeat encodes a stretch of glutamic acid (E) residues. Two of the IMAGE clones contained six GGA repeats, whereas the SEREX-

derived clone and three IMAGE clones contained seven repeats. A second set of two GGA repeats is present seven bases downstream, encoding two glycine (G) residues. No variation in the length of this second repeat was seen among the SEREX-derived clone and the IMAGE clones. Both repeats are present within the C-terminal region that has homology to the MAGE superfamily proteins, however the presence of these repeats is a unique feature of *MAGE-F1*. Two other sequence variations were noted. The IMAGE clone sequences yielded either a G or T at position 372, encoding alanine (A) or serine (S) residues, respectively, while position 886 yielded either T or G, encoding alanine (A) or glutamic Acid (E). These differences likely represent mutations or single base poly-morphisms.

3.2. *MAGE-F1* is located on chromosome 3, with the entire coding sequence included in a single exon

The *MAGE-F1* cDNA includes the STS WI-13826 (Fig. 1). As reported in the original unigene cluster Hs.84461, this marker has been mapped to chromosome 3 on the radiation hybrid map. However, shortly following our

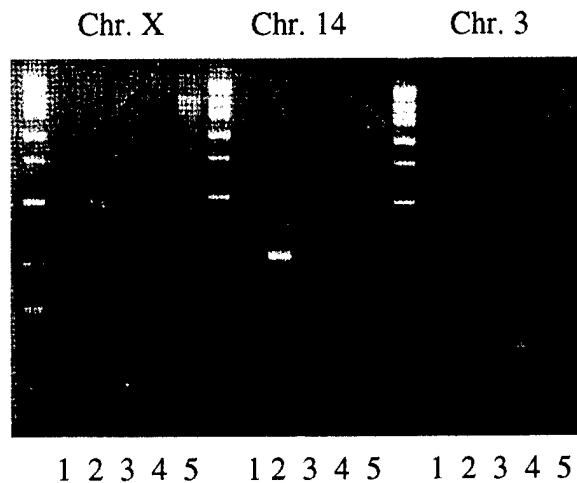


Fig. 3. Mapping WI-13826 to human Chromosome 3 by PCR from rodent/human cell hybrids containing human chromosomes X, 14 or 3 as indicated at the top of the figure (Coriell Scientific). Each of the three genomic DNA templates was amplified with primers for the following Sequence Tagged Sites: 1, DXS995 (X chromosome); 2, D14S1066 (chromosome 14); 3, D3S1580 (chromosome 3); 4, D3S1553 (chromosome 3); 5, WI13826 (*MAGE-F1* locus) using standard conditions presented in Section 2 and GenBank.

cloning of *MAGE-F1*, Hs.84461 was retired and replaced with a new cluster, Hs.75621. The new cluster includes mapping information from chromosomes 14, 3 and X. To clarify which chromosome encodes *MAGE-F1*, PCR analysis was carried out using human chromosome-specific DNA from rodent/human hybrid cell lines (Coriell Cell Repositories). Using genomic DNA templates derived from lines harboring a single human chromosome (3, 14 or X), STSs flanking the intervals reported in Hs.75621 were amplified. For each chromosome-specific template, control STS primers amplified the expected fragment from the appropriate DNA sample, indicating that the three DNA samples contained the expected human chromosomes (Fig. 3). Amplification of WI-13826 was observed only with template containing human chromosome 3 DNA. In parallel experiments, the entire coding sequence of *MAGE-F1* could be amplified using genomic DNA as a template, therefore the *MAGE-F1* coding region is included in a single exon (data not shown).

Most *MAGE* family genes are clustered in four different loci, designated A through D, on the X chromosome (Lucas et al., 1999). *NDN*, a distant *MAGE* relative, is encoded on chromosome 15 (Jay et al., 1997; Nakada et al., 1998; Watrin et al., 1997). Our finding that *MAGE-F1* is located on chromosome 3 identifies a sixth locus for a *MAGE* superfamily gene. Searches of the Unigene database and websites of the major genome mapping centers did not reveal evidence of other *MAGE* superfamily genes on chromosome 3.

3.3. Expression of *MAGE-F1*

Most members of the *MAGE* superfamily exhibit a

cancer/testis or 'CT' expression pattern in which expression in the adult is restricted to testis, placenta and a subset of tumors (Van den Eynde and Boon, 1997; van Baren et al., 1999). However, *MAGE-D1* and *MAGE-D2/JCL-1* have recently been shown to have more ubiquitous expression patterns in normal adult tissues, and *NDN* can be detected in many tissues by RT-PCR (Lucas et al., 1999; Pold et al., 1999; Nakada et al., 1998). We assessed expression of *MAGE-F1* using an MTE membrane from Clontech. This membrane contains normalized mRNA samples from 76 adult and fetal human tissues. Upon hybridization with a radiolabeled *MAGE-F1* cDNA probe, a clear signal was obtained from all tissues, though not from negative controls such as human Cot1 DNA or yeast DNA (Fig. 4A). Long exposures revealed a faint signal from the spot corresponding to 500 ng of human genomic DNA. Sequential stripping and re-hybridization with control probes for *ubiquitin* and the CT antigen *NY-ESO-1* produced the expected ubiquitous and testis-restricted hybridization patterns, respectively

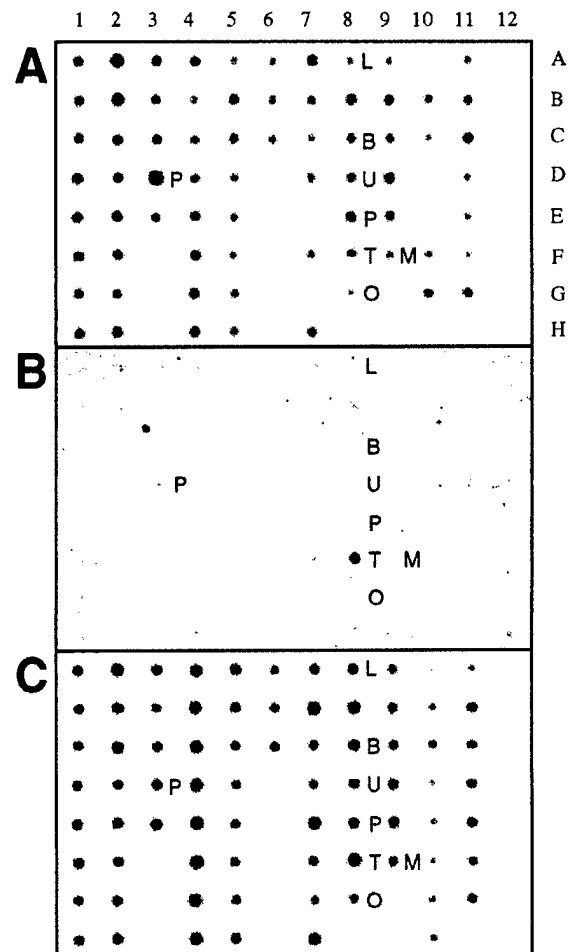


Fig. 4. Expression of *MAGE-F1* and *NY-ESO-1* mRNA in multiple human tissues. MTE membranes were hybridized to radiolabeled probes corresponding to (A) *MAGE-F1*, (B) *NY-ESO-1* and (C) *Ubiquitin*. A complete key listing all tissues is printed in Section 2. P, pituitary; L, lung; B, Bladder; U, uterus; P, prostate; T, testis; O, ovary; M, mammary gland.

(Chen et al., 1997) (Fig. 4B,C). Compared to the *ubiquitin* control probe, there was a clear elevation of *MAGE-F1* signal in mRNA derived from the pituitary. Thus, it appears that *MAGE-F1* has a ubiquitous expression pattern similar to that reported for *MAGE* genes encoded at the D locus.

As noted above, *MAGE-F1* protein also shows a high level of homology to the uncharacterized gene product, KIAA1114. While there are currently no published data concerning the expression pattern of KIAA1114, the number of ESTs reported for this gene, 69, is in the range found with other *MAGE* genes that are ubiquitously expressed (238 ESTs for *MAGE-D1*, 283 ESTs for *MAGE-D2/JCL-1*, 59 for *NDN* and 77 for *MAGE-F1*; Table 2). In contrast, *MAGE* genes exhibiting CT expression patterns have relatively few ESTs, ranging from 0 (*MAGE B4*) to 12 (*MAGE-A10*). Thus, when both protein sequence homology and expression patterns are considered, it appears that *MAGE-F1* is most closely related to a subset of *MAGE* genes that are ubiquitously expressed in normal adult tissues. This group includes *MAGE-D1*, *MAGE-D2/JCL-1* and *NDN*, with the likely addition of several isoforms of *KIAA1114*. Interestingly, all mapped *MAGE* genes exhibiting a CT expression pattern are located at the A, B or C loci on the X chromosome, whereas those mapped genes with ubiquitous expression are located at or near the D locus on Xp11 (*MAGE-D1*, *MAGE-D2/JCL-1*), or on chromosomes 15 (*NDN*) or 3 (*MAGE-F1*). While *KIAA1114* has not been formally mapped, it is included on a genomic clone (AL049732) from Xp11. In addition, KIAA1114 appears to be variably spliced, and contains 2483 bases of nucleotide sequence identity with the *Trophinin* gene on Xp11 (Pack et al., 1997). Therefore, it is likely that *Trophinin* is a splice variant of *KIAA1114*, placing *KIAA1114* at the D locus on Xp11 as well.

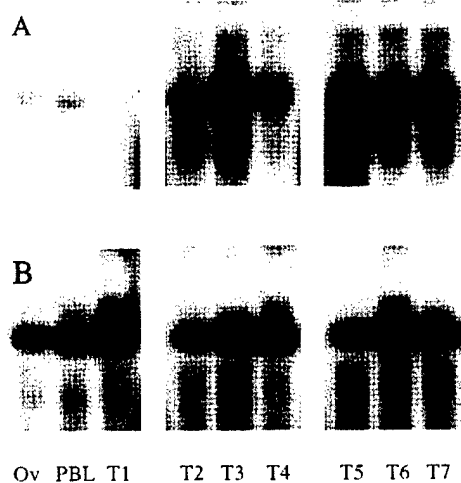


Fig. 5. Northern blots showing expression of *MAGE-F1* mRNA in normal ovary (Ov), Peripheral Blood Lymphocytes (PBL) and seven stage III/IV serous ovarian tumors, (T1–T7). Panel A was probed with a radiolabeled PCR fragment for the *MAGE-F1* coding sequence. Panel B was probed with a radiolabeled *Ubiquitin* cDNA to confirm equivalent RNA loading.

As *MAGE-F1* was cloned from an ovarian tumor cDNA library, we examined *MAGE-F1* expression in multiple ovarian tumor specimens. Northern blots were performed using RNA derived from seven late-stage serous ovarian tumors, normal whole ovary and peripheral blood lymphocytes (PBL). A single 1.7 kb band, corresponding to *MAGE-F1*, was detected in normal ovary, PBL and six tumors (Fig. 5). One tumor showed negligible *MAGE-F1* expression. The six *MAGE-F1* expressing tumors all showed more intense *MAGE-F1* hybridization than control whole ovary or PBL. Stripping and re-hybridizing the blot with a *Ubiquitin* cDNA probe demonstrated equivalent RNA loading in all lanes.

Real-time reverse-transcriptase PCR (RT-PCR) was used to examine *MAGE-F1* expression in a larger number of normal and diseased ovarian surgical specimens. Consistent with the Northern blots, *MAGE-F1* mRNA was expressed in 17/17 normal whole ovaries, 7/7 benign ovarian tumors, 2/2 borderline ovarian tumors, 21/21 stage III/IV ovarian tumors and 5/5 metastatic ovarian tumors. On average, the abundance of *MAGE-F1* mRNA was similar among normal and diseased ovarian specimens (Fig. 6). RT-PCR also revealed expression of *MAGE-F1* in ovarian, cervical, breast and melanoma tumor cell lines.

3.4. Potential immunological crossreactivity of *MAGE-F1* with other *MAGE* proteins

Several *MAGE* genes with CT expression patterns are currently in use as target antigens for immunotherapy (Mackensen et al., 2000; Weber et al., 1999; Reynolds et al., 1997; 1998). Lucas et al. raised the possibility that cross-reactivity between CT class *MAGE* proteins and the ubiquitously expressed *MAGE-D2/JCL-1* could result in autoimmune destruction of normal host tissues (Lucas et al., 1999).

Expression of *MAGE-F1* and β -actin in ovarian tissues and cell lines.

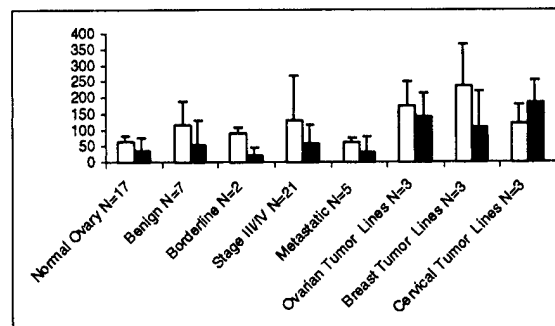


Fig. 6. Expression of *MAGE-F1* (open bars) and β -actin (filled bars) mRNA in normal ovary, staged ovarian tumors and tumor cell lines assessed by real-time PCR. All PCR amplifications were performed in duplicate and the presence of single reaction products of the expected size was verified by agarose gel electrophoresis. Values are given in arbitrary units relative to the amplification standard S31iii125 from the control PBL derived cDNA dilution series. Error bars represent one standard deviation of the mean. N, number of independent tissue samples analyzed.

Table 3

Known T cell epitopes encoded by CT class MAGE superfamily genes and corresponding HLA restriction elements^a

MAGE gene(s)	Sequence	Restriction element
A1	SAFPTINF	Cw2
A1	SLFRAVITK	A3
A1	NYKHCFPEI	A24
A1	EADPTGHSY	A1
A1	EVYDGREHSA	A28
A1	SAYGEPRKL	Cw3, Cw16
A1	DPARYEFLW	B53
A2	YLQLVFGIEV	A2
A3	TSYVKVLHMHVKISG	DR11
A3	KVAELVHFL	A2
A3	LLKYRAREPVTKEAE	DR13
A3	EVDPIGHLY	B44, A1
A3	IMPKAGLLI	A24
A3a	GVYDGREHTV	A2
A3, A12	FLWGPRLV	A2
A6	MKISGGPR	A34
A10	GLYDGMHL	A2
B1, B2	FLWGPRAYA	A2

^a Adapted from Lucas et al., 1999 and references therein.

However, they found that none of the known T-cell epitopes from CT class MAGE proteins are present in MAGE-D2/JCL1. Our finding that MAGE-F1 is also expressed in most normal tissues raised the same concern about potential cross-reactivity. However, as with *MAGE-D2/JCL-1*, the *MAGE-F1* sequence does not encode any of the known T cell epitopes from CT class MAGE proteins (see Table 3 for peptides analyzed). Indeed, the MAGE-F1 peptide with the closest match to a known T-cell epitope contained four substitutions in a nine amino acid stretch. Thus, cross-reactivity between therapeutic MAGE peptides and MAGE-F1 appears unlikely. Nevertheless, epitope spreading leading to cross-reactivity between a CT class MAGE protein and one of the ubiquitously expressed MAGE proteins remains a possibility.

3.5. Possible function of MAGE proteins

The observation that several members of the *MAGE* superfamily are ubiquitously expressed suggests that they may play a functional role in normal cellular physiology. Several recent publications have shown that necdin, the murine homologue to the human NDN protein, binds p53 and E2F, as well as the viral proteins E1A and large T antigen (Taniura et al., 1998, 1999). It has been proposed that Necdin may function in an Rb-like manner in post-mitotic neurons, contributing to growth arrest by repressing E2F-dependent transcription (Taniura et al., 1998; Yoshikawa, 2000). Furthermore, over-expression of Necdin has growth inhibitory effects in cultured cell lines (Hayashi et al., 1995). Deletion analysis has shown that the p53 binding domain of Necdin corresponds to the most highly conserved region of MAGE superfamily proteins (Taniura et al., 1999). This

suggests that MAGE-F1 and other MAGE proteins may also be involved in the control of cell growth and viability in normal and neoplastic cells. If so, the aberrant expression of MAGE proteins in a wide variety of human cancers may not be an inadvertent consequence of altered DNA methylation, as is currently assumed, but rather may be a contributing factor to the transformed phenotype by altering cell cycle regulation or the response to DNA damage.

4. Conclusions

- We report the cloning and sequencing of MAGE-F1 cDNA, a novel member of the MAGE gene superfamily.
- MAGE-F1 is ubiquitously expressed in normal tissues, as well as melanoma, leukemia, ovarian and cervical tumors and cell lines.
- MAGE-F1 is encoded on chromosome 3, a novel locus for MAGE family genes.
- No known MAGE-family T-cell epitopes are found in the MAGE-F1 coding sequence.

Acknowledgements

We would like to thank Dr Charles Drescher and the Marsha Rivkin Center Working Group for valuable scientific discussions and support, and Patty Theiss for help in the preparation of this manuscript. This work was supported by grants from the U.S. Department of Defense (OC970002), the Marsha Rivkin Center for Ovarian Cancer Research, the Morrison Trust, the Wilkins Fund, the M.J. Murdock Charitable Trust and the William Randolph Hearst Foundation.

References

- Chen, Y.T., Scanlan, M.J., Sahin, U., Tureci, O., Gure, A.O., Tsang, S., Williamson, B., Stockert, E., Pfreundschuh, M., Old, L.J., 1997. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *Proc. Natl. Acad. Sci. USA* 94, 1914–1918.
- De Plaen, E., Arden, K., Traversari, C., Gaforio, J.J., Szikora, J.P., De Smet, C., Brasseur, F., van der Bruggen, P., Lethe, B., Lurquin, C., 1994. Structure, chromosomal localization, and expression of 12 genes of the MAGE family. *Immunogenetics* 40, 360–369.
- De Smet, C., Lurquin, C., Lethe, B., Martelange, V., Boon, T., 1999. DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Mol. Cell Biol.* 19, 7327–7335.
- Gure, A.O., Stockert, E., Arden, K.C., Boyer, A.D., Viars, C.S., Scanlan, M.J., Old, L.J., Chen, Y., 2000. CT10: A new cancer-testis (CT) antigen homologous to CT7 and the MAGE family, identified by representational-difference analysis. *Int. J. Cancer* 85, 726–732.
- Hayashi, Y., Matsuyama, K., Takagi, K., Sugiura, H., Yoshikawa, K., 1995. Arrest of cell growth by necdin, a nuclear protein expressed in post-mitotic neurons. *Biochem. Biophys. Res. Commun.* 213, 317–324.
- Henikoff, S., Henikoff, J.G., 1994. Protein family classification based on searching a database of blocks. *Genomics* 19, 97–107.

- Huang, L.Q., Brasseur, F., Serrano, A., De Plaen, E., van der Bruggen, P., Boon, T., Van Pel, A., 1999. Cytolytic T lymphocytes recognize an antigen encoded by MAGE-A10 on a human melanoma. *J. Immunol.* 162, 6849–6854.
- Jay, P., Rougeulle, C., Massacrier, A., Moncla, A., Mattei, M.G., Malzac, P., Roeckel, N., Taviaux, S., Lefranc, J.L., Cau, P., Berta, P., Lalande, M., Muscatelli, F., 1997. The human necdin gene, *NDN*, is maternally imprinted and located in the Prader-Willi syndrome chromosomal region. *Nat. Genet.* 17, 357–361.
- Lucas, S., De Smet, C., Arden, K.C., Viars, C.S., Lethe, B., Lurquin, C., Boon, T., 1998. Identification of a new MAGE gene with tumor-specific expression by representational difference analysis. *Cancer Res.* 58, 743–752.
- Lucas, S., Brasseur, F., Boon, T., 1999. A new MAGE gene with ubiquitous expression does not code for known MAGE antigens recognized by T cells. *Cancer Res.* 59, 4100–4103.
- Lucas, S., De Plaen, E., Boon, T., 2000. MAGE-B5, MAGE-B6, MAGE-C2, and MAGE-C3: Four new members of the MAGE family with tumor-specific expression. *Int. J. Cancer* 87, 55–60.
- Lurquin, C., De Smet, C., Brasseur, F., Muscatelli, F., Martelange, V., De Plaen, E., Brasseur, R., Monaco, A.P., Boon, T., 1997. Two members of the human MAGEB gene family located in Xp21.3 are expressed in tumors of various histological origins. *Genomics* 46, 397–408.
- Mackensen, A., Herbst, B., Chen, J.L., Kohler, G., Noppen, C., Herr, W., Spagnoli, G.C., Cerundolo, V., Lindermann, A., 2000. Phase I study in melanoma patients of a vaccine with peptide-pulsed dendritic cells generated in vitro from CD34(+) hematopoietic progenitor cells. *Int. J. Cancer* 86(3), 385–392.
- Nakada, Y., Taniura, H., Uetsuki, T., Inazawa, J., Yoshikawa, K., 1998. The human chromosomal gene for necdin, a neuronal growth suppressor, in the Prader-Willi syndrome deletion region. *Gene* 213, 65–72.
- Pack, S.D., Tanigami, A., Ledbetter, D.H., Sato, T., Fukuda, M.N., 1997. Assignment of trophoblast/endometrial epithelium cell adhesion molecule trophinin gene *TRO* to human chromosome bands Xp11.22→p11.21 by in situ hybridization. *Cytogenet. Cell Genet.* 79, 123–124.
- Pold, M., Zhou, J., Chen, G.L., Hall, J.M., Vescio, R.A., Berenson, J.R., 1999. Identification of a new, unorthodox member of the MAGE gene family. *Genomics* 59(2), 161–167.
- Reynolds, S.R., Oratz, R., Shapiro, R.L., Hao, P., Yun, Z., Fotino, M., Vukmanovic, S., Bystry, J.C., 1997. Stimulation of CD8 + T cell responses to MAGE-3 and Melan A/MART-1 by immunization to a polyvalent melanoma vaccine. *Int. J. Cancer* 72, 972–976.
- Reynolds, S.R., Celis, E., Sette, A., Oratz, R., Shapiro, R.L., Johnston, D., Fotino, M., Bystry, J.C., 1998. HLA-independent heterogeneity of CD8 + T cell responses to MAGE-3, Melan A/MART-1, gp100, tyrosinase, MC1R, and TRP-2 in vaccine-treated melanoma patients. *J. Immunol.* 161, 6970–6976.
- Sahin, U., Tureci, O., Schmitt, H., Cochlovius, B., Johannes, T., Schmits, R., Stenner, F., Luo, G., Schobert, I., Pfreundschuh, M., 1995. Human neoplasms elicit multiple specific immune responses in the autologous host. *Proc. Natl. Acad. Sci. USA* 92, 11810–11813.
- Serrano, A., Garcia, A., Abril, E., Garrido, F., Ruiz-Cabello, F., 1996. Methylated CpG points identified within MAGE-1 promoter are involved in gene regression. *Int. J. Cancer* 68(4), 464–470.
- Taniura, H., Taniguchi, N., Hara, M., Yoshikawa, K., 1998. Necdin, a postmitotic neuron-specific growth suppressor, interacts with viral transforming proteins and cellular transcription factor E2F1. *J. Biol. Chem.* 273, 720–728.
- Taniura, H., Matsumoto, K., Yoshikawa, K., 1999. Physical and functional interactions of neuronal growth suppressor necdin with p53. *J. Biol. Chem.* 274, 16242–16248.
- Tureci, O., Sahin, U., Pfreundschuh, M., 1997. Serological analysis of human tumor antigens: molecular definition and implications. *Mol. Med. Today* 3, 342–349.
- Tureci, O., Sahin, U., Zwick, C., Neumann, F., Pfreundschuh, M., 1999. Exploitation of the antibody repertoire of cancer patients for the identification of human tumor antigens. *Hybridoma* 18, 23–28.
- van Baren, N., Brasseur, F., Godelaine, D., Hames, G., Ferrant, A., Lehmann, F., Andre, M., Ravoet, C., Doyen, C., Spagnoli, G.C., Bakkus, M., Thielemans, K., Boon, T., 1999. Genes encoding tumor-specific antigens are expressed in human myeloma cells. *Blood* 94, 1156–1164.
- Van den Eynde, B.J., Boon, T., 1997. Tumor antigens recognized by T lymphocytes. *Int. J. Clin. Lab Res.* 27, 81–86.
- van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den, E.B., Knuth, A., Boon, T., 1991. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* 254, 1643–1647.
- Watrif, F., Roeckel, N., Lacroix, L., Mignon, C., Mattei, M.G., Distèche, C., Muscatelli, F., 1997. The mouse Necdin gene is expressed from the paternal allele only and lies in the 7C region of the mouse chromosome 7, a region of conserved synteny to the human Prader-Willi syndrome region. *Eur. J. Hum. Genet.* 5, 324–332.
- Weber, J.S., Hua, F.L., Spears, L., Marty, V., Kuniyoshi, C., Celis, E., 1999. A phase I trial of an HLA-A1 restricted MAGE-3 epitope peptide with incomplete Freund's adjuvant in patients with resected high-risk melanoma. *J. Immunother.* 22, 431–440.
- Yoshikawa, K., 2000. Cell cycle regulators in neural stem cells and post-mitotic neurons [In Process Citation]. *Neurosci. Res.* 37, 1–14.
- Zorn, E., Hercend, T., 1999. A MAGE-6-encoded peptide is recognized by expanded lymphocytes infiltrating a spontaneously regressing human primary melanoma lesion. *Eur. J. Immunol.* 29(2), 602–607.

Appendix G

Core: Related Publications

- Pepe, MS, Longton G., Anderson, GL., Schummer, MS. Selecting Differentially Expressed Genes from Microarray Experiments. (In press, Biometrics).
- McIntosh MW, Urban N. A Parametric Empirical Bayes Method for Cancer Screening Using Longitudinal Observations of a Tumor Marker. (In press, Biostatistics).
- McIntosh MW, Urban N, Karlan B. Generating Longitudinal Cancer Screening Algorithms for Novel Tumor Markers. (In press, Cancer Epidemiology Biomarkers and Prevention)

Selecting Differentially Expressed Genes from Microarray Experiments

Margaret Sullivan Pepe,^{1,2} Gary Longton,² Garnet Anderson,² and Michel Schummer³

¹Department of Biostatistics, University of Washington

Seattle, Washington 98195-7232, U.S.A

²Division of Public Health Sciences, Fred Hutchinson Cancer Research Center

Seattle, Washington 98109-1024, U.S.A.

³Institute for Systems Biology, Seattle, Washington 98105-6099, U.S.A.

**email: mspepe@u.washington.edu*

SUMMARY. High throughput technologies, such as gene expression arrays and protein mass spectrometry, allow one to simultaneously evaluate thousands of potential biomarkers that distinguish different tissue types. Of particular interest here are cancer versus normal organ tissues. We consider statistical methods to rank genes (or proteins) in regards to differential expression between tissues. Various statistical measures are considered and we argue that two measures related to the Receiver Operating Characteristic Curve are particularly suitable for this purpose. We also propose that sampling variability in the gene rankings be quantified and suggest using the 'selection probability function,' the probability distribution of rankings for each gene. This is estimated via the bootstrap. A real data set derived from gene expression arrays of 23 normal and 30 ovarian cancer tissues are analyzed. Simulation studies are also used to assess the relative performance of different statistical gene ranking measures and our

quantification of sampling variability. Our approach leads naturally to a procedure for sample size calculations appropriate for exploratory studies that seek to identify differentially expressed genes.

KEY WORDS: Classification; Discrimination; Exploratory analysis; Genomics; Prediction; Proteomics; ROC curves.

1. Introduction

The development of microarrays that provide simultaneous evaluation of mRNA expression levels for thousands of genes is one of the most exciting new advances in modern medical research. It promises to identify disease at its most basic biological level, namely at that of the genes. The implications for medicine are enormous (The Chipping Forecast, 1999). Insights into genetic alterations caused by disease can lead to new therapeutic strategies. Genetic alterations that precede disease can be targets for disease prevention strategies. The research community can expect insights into the etiology of disease and pathways involved in its progression, that will surely revolutionize medical practice.

There are new statistical challenges posed by data from microarray experiments, due primarily to the exploratory nature of experiments and the huge numbers of genes under investigation. It must also be recognized that different sorts of questions are addressed with microarray experiments and that the appropriate statistical approach depends, of course, on the question of interest (Dudoit et al, 2000a). Categories of objectives

pertaining to experiments that include multiple tissue types (e.g. cancer versus non-cancer tissue) include: (i) selection of genes that are differentially expressed in different known classes of tissue; (ii) identification of a minimal combination of genes that provides discrimination between known tissue types; (iii) identification of groups of genes whose expression levels are correlated; and (iv) new classifications of tissue types defined by genes whose expression levels are related. Statistical techniques such as regression methods and discriminant analyses have been adapted for (ii) (Dudoit et al, 2000a), whereas clustering techniques are more appropriate for (iii) and (iv) (Tibshirani et al, 2000, Hastie et al, 2000, Lazzeroni and Owen, 2000, Van Der Laan and Bryan, 2000). In this paper we consider statistical methods for objective (i), which, at first glance, seems to be the most straight-forward.

The particular application that motivated our work concerns the search for biomarkers of ovarian cancer that could be used in population screening. Ovarian tissue from 30 subjects with cancer and 23 subjects without cancer were analyzed for mRNA expression using glass arrays spotted for 1536 gene clones. The data, Y_{ig} , for the g^{th} gene clone in the i^{th} tissue sample, is a measure of the mRNA expression of the g^{th} gene in that tissue relative to a control tissue, with a common control employed for all experiments. We refer to Dudoit et al, 2000b and Newton et al, 2000 for a summary of this technology and a technical explanation for how Y_{ig} is calculated. Using standard terminology for these experiments, Y_{ig} is the ratio of the intensities of the red to green fluorescent dyes, where green dye is used for the common control and red is used for the experimental tissue.

The scientific objective is to identify genes that are differentially expressed in ovarian cancer tissue compared with normal ovarian tissue. Ovarian tissue cannot of course be used directly for population screening. However, if a gene is found that is expressed differentially in cancer, then the corresponding protein product (or an antibody to it) may be detectable in blood or urine and could be the basis for a population screening test (Pepe et al 2001). Scientists are more interested in identifying genes that are over-expressed rather than under-expressed in cancer, because detecting the presence of a new aberrant protein in blood is a potentially easier task than detecting the absence of a normal protein, particularly if that protein is also produced by other tissues in the body.

There are many genes over-expressed in cancer tissue that cannot lead to screening markers. For example, genes that relate simply to inflammation or growth are not candidates because those processes also occur naturally in the body. Clinical assays for some gene products may be too difficult to develop for technical reasons. Therefore we need to select a sizeable number of over-expressed genes in order to arrive at a subset that might have potential for screening. For the initial selection, we will include multiple genes that are redundant in the sense that they identify the same cancer samples so that if one gene proves useless for biomarker development we can still pursue another that could identify those same cancers.

The experimental data are used to rank candidate genes according to some statistical measure characterizing differential expression. In section 2 we discuss the choice of

statistical measure. A method for quantifying the degree of confidence in the ranking of a gene provided by the data is proposed in section 3. This acknowledges the finite number of tissues examined, variability across tissues and the large number of genes investigated, all of which contribute to uncertainty in the ranking of the genes. Application to the ovarian cancer data in section 4 illustrates the approach. Recommendations for computing sample sizes in these exploratory studies are provided in section 5. Some further remarks about experimental design are made in section 6. We close with some thoughts on further extensions of our proposed methods.

2. Characterizing Interesting Differential Expression

2.1 Measures of Discrimination

At each gene, data are available for n_D cancer tissues and n_C normal tissues.

$$\left\{ \begin{array}{l} Y_{gi}^D, i = 1, \dots, n_D \\ Y_{gj}^C, j = 1, \dots, n_C \end{array} \right\}.$$

To say that there is differential expression at gene g is to say that the distribution of Y_g^D is different from that for Y_g^C . What sorts of differences are of particular interest? Figure 1 displays some hypothetical distributions that we use for discussion. Without loss of generality Y_g^C has a standard normal distribution or the data can be transformed to achieve this.

The ideal situation is represented in the top panel where there is almost complete separation between the distributions. In this case the relative expression level of gene g is an ideal candidate marker for cancer because the values are completely different in cancer tissue from those in normal tissue. There is a threshold value that allows one to classify cancer versus normal tissue with almost 100% accuracy.

Consider now settings where the distributions overlap. We contend that for cancer screening, the separation in panel II is of more practical interest than that in panel III. The marker clearly distinguishes a subset of cancers from normals in II, whereas in panel III marker values for cancer tissues are entirely within the range of those for non-cancer tissues. Looking ahead to population screening and assuming that gene expression translates roughly into protein expression, in panel II there is a threshold for the screening test that provides detection of about 30% of cancers while falsely identifying only 1% of non-cancers as screen positive. In screening it is important to keep false positive rates extremely low because even a small false positive rate translates into large numbers of people being subjected to diagnostic procedures that are costly and invasive. Using a similar threshold in panel III corresponding to the 1% false positive rate, detects only 2% of cancers because the distributions overlap over the whole normative range.

We suggest that statistical measures of discrimination between the distribution of Y_g^D and Y_g^C focus on separation at and beyond upper quantiles of the normative range. Figure 2 shows receiver operating characteristic (ROC) curves that characterize separations between distributions. Each point on the ROC curve, $(t, \text{ROC}(t))$, corresponds to a

different threshold u , and by definition $t = P[Y_g^C \leq u]$, and $ROC(t) = P[Y_g^D \geq u]$. The ROC curve can be thought of as a plot of the true versus false positive rates associated with all possible thresholds for classifying a tissue as cancerous based on the relative expression level Y_g (Pepe, 2000). Because low values of t correspond to high quantiles of Y_g^C , our suggestion is to focus on the ROC curve at low values of t .

Two summary measures of discrimination that we propose are:

$$ROC(t_0) = P[Y_g^D \geq y^C(1-t_0)]$$

and

$$pAUC(t_0) = \int_0^{t_0} ROC(t) dt.$$

where t_0 is some small false positive rate and $y^C(1-t_0)$ is the quantile in the upper tail of the normative range corresponding to t_0 . The measure $ROC(t_0)$ is easily conceived of by non-statisticians, as the proportion of cancer tissues with expression levels above the $(1-t_0)$ quantile of normal tissues. The partial area under the curve, $pAUC(t_0)$, in effect averages this proportion across values of $t < t_0$ (McClish, 1989). If two curves have the same value of $ROC(t_0)$, the curve with larger $pAUC(t_0)$ would indicate better

separation at that gene because for some values of $t < t_0$, $\text{ROC}(t)$ must be higher for that gene.

The $\text{ROC}(t_0)$ or $\text{pAUC}(t_0)$ statistic calculated for the three settings of Figure 1, ranks biomarker II better than biomarker III for small values of t_0 ($t_0 \bullet 0.10$). On the other hand, other classic measures of discrimination such as the two-sample t-statistic or the Mann-Whitney U statistic (equivalently the Wilcoxon statistic) rank biomarker III better than biomarker II. We regard this as a weakness of those statistics for our application. We also see from Figure 2 that all of these statistics rank biomarker I as the best, regardless of t_0 , and indeed any reasonable statistic should because biomarker I is almost perfect.

How should one choose t_0 ? Ideally the choice of t_0 will depend on false positive rates that are acceptable in practice, and t_0 could be chosen as the maximally acceptable one. The magnitudes of false positive rates that are acceptable will vary with the application since they depend on the costs and consequences of the errors. Very small t_0 are in general required for cancer screening. However, with small numbers of tissue samples, estimates of $\text{pAUC}(t_0)$ or $\text{ROC}(t_0)$ at very small t_0 will not be possible. Thus in our application we chose t_0 to be small, but large enough that a viable estimate of the $\text{pAUC}(t_0)$ can be calculated. Further research into appropriate choices for t_0 would be worthwhile.

We suggest that empirical estimates of $ROC(t_0)$ and $pAUC(t_0)$ be used to rank genes for differentiated expression in cancer versus normal tissue. Other measures of discrimination that we calculate are (i) Zstat, the standardized difference in means, i.e., the two-sample t-statistic and (ii) AUC, the area under the entire ROC curve

$$AUC = \int_0^1 ROC(t)dt.$$

Interestingly the empirical AUC is equivalent to the numerator of the Mann-Whitney U-statistic, $\sum_i \sum_j I[Y_{gi}^D \geq Y_{gj}^C]/n_D n_C$, for comparing the distribution of Y_g^D and Y_g^C and can be interpreted as an estimate of $P[Y_g^D \geq Y_g^C]$ (Bamber, 1975). Each of $ROC(t_0)$, $pAUC(t_0)$ and AUC are distribution free rank statistics whereas Zstat depends on the underlying probability distributions for Y_g^D and Y_g^C .

2.2 Illustration

As a small illustration we evaluated the first 100 genes in our ovarian cancer dataset. Table 1 displays the top 10 ranking genes in order when ranked according to the different statistical measures. To a large extent the same genes were identified by all discrimination measures, although the order of ranking differed. Consider, however, genes 5 and 97 for which raw data and ROC curves are displayed in Figure 3. The Mann-Whitney U-statistic (AUC) ranked these genes very similarly, as the 6th and 8th,

respectively. On the other hand, the pAUC statistic ranked them quite differently as the 3rd and 31st ranking genes, respectively. The raw data and the ROC curves indicate that indeed for gene 5 more of the cancer tissues are above the bulk of the normative range than is the case for gene 40. The pAUC statistic picks up on this fact and gives it a far higher rank than it gives gene 97. It suggests to these authors that gene 5 should receive higher priority for biomarker development than gene 97.

Insert Table 1

3. Assessing Variability

3.1 The probability of gene selection.

The relative rankings of genes is the primary outcome of the study. However, the rankings are subject to sampling variability. How should this variability be acknowledged? Standard errors or p-values don't seem to be directly relevant to the task because the objectives of the study are neither to estimate parameters nor to test hypotheses. Rather, the task is to rank genes and to select the top genes for further study. Therefore we propose the following quantity to quantify our degree of confidence in choosing the g^{th} gene among the top k .

$$\begin{aligned} P_g(k) &= P[\text{gene } g \text{ ranked in the top } k] \\ &= P[\text{Rank}(g) \leq k] \end{aligned}$$

The value of $P_g(k)$ may be of particular interest for k equal to a predetermined number of genes to be selected (10 in the small illustration). However, the whole survivor function can be considered, $\{P_g(k), k \geq 1\}$, and this gives a more full description of sampling variability in the ranking. Various factors contribute to the variability in Rank(g): (i) the number of cancer tissues and normal tissues studied, n_D and n_C ; (ii) the extent and type of differential expression of the g^{th} gene; (iii) the number of genes in the selection pool, which we denote by N ; (iv) the differential expression of genes other than the g^{th} gene; and not least, (v) the algorithm used to rank genes. The quantity, $P_g(k)$, will be affected by all of these factors.

Intuitively, as sample sizes increase, the $P_g(k)$ function will tend to 0 or 1 according to whether the true discriminating measure for the g^{th} gene ranks below k or not. Genes that in truth are very highly discriminatory will certainly have high ranks even in experiments with small sample sizes and $P_g(k)$ will be close to 1. This may be reduced by chance if there is a large number of competing genes and in particular if a substantial number of competing genes also exhibit differential expression. Observe that at the opposite extreme, if no genes are differentially expressed, then $P_g(k) = k/N$.

The selection probabilities, $P_g(k)$, as we call them, can be estimated by the bootstrap with the resampling unit being at the tissue level. Thus, when a tissue is included in the bootstrap sample, the entire vector of data relating to all genes for that tissue is entered

into the bootstrapped dataset, and genes are ranked within the dataset according to the statistical measure chosen. The bootstrapping therefore acknowledges the complex correlations between genes.

All of our statistical measures but Zstat are rank statistics. Tied data points influence the distribution of rank statistics and we note that tied data points ensue with simple resampling of observed data. However, real data, such as the original dataset, do not have ties because Y_g is measured on a continuous scale. Thus, we modified the bootstrapping to randomly break ties by adding miniscule random noise (jitter) to the expression levels. This was done in an effort to make the bootstrap distribution of the rank statistics more reflective of the actual distribution across different realizations of the experiment.

3.2 Illustration

Returning to the small illustration described earlier, Table 1 shows $P_g(10)$ based on 200 bootstrapped samples for each gene ranked in the top 10. Thus if the strategy of the experiment is to select the top 10 genes ranked on the basis of ROC(0.1), we are highly confident ($P_g(10) > 90\%$) about the selection of genes 93, 76, 65 and 42. However, we estimate that, due to sampling variability, genes 35, 23 and 52 have $\leq 60\%$ chance of ranking in the top 10 if the experiment were repeated.

A comparison of the two estimators of $P_g(10)$, with and without jitter in the bootstrap sample, suggested that they are quite similar. That is, we arrive at the same conclusions about $P_g(10)$ for the rank-based measures if the data are jittered or not. Thus tied datapoints in the bootstrap samples do not appear to affect the $P_g(10)$ estimates substantially.

3.3 Simulation

As a simple example we simulated data on 2000 genes for equal numbers of cancers and normal tissues. Of the 2000 genes, 95% were configured to be non-informative in the sense that Y_g^D and Y_g^C had the same distributions, namely standard normal. For 100 genes the distributions were normal with mean 1 and standard deviation 2 for cancer tissues and standard normal for non-cancer tissues. For an informative gene therefore, the area under the corresponding ROC curve was $\Phi((1-0)/\sqrt{2^2+1^2}) = 0.67$ (Reiser and Guttman, 1986). Data for different genes were generated independently. We set the number of genes to be selected at $k = 100$. Table 2 panel A shows the proportions of informative markers selected averaged across 100 simulation studies. That is, it shows $P[\text{Rank}(g) \leq k \mid g \text{ is an informative gene}]$.

Insert table 2

The results suggest that the top 100 genes consist primarily of informative genes even with relatively small sample sizes. An informative gene has a 68% chance of being in the top 100 ranked on the basis of the $pAUC(0.2)$ statistic when 30 samples are analyzed, 15 cancer and 15 normal tissues. The chance reaches 91% when a total of 35 cancer and 35 normals are evaluated.

In this particular example, (setting A of Table 2), the $pAUC$ statistic was most effective at selecting informative genes. Interestingly it outperformed the full area under the curve the AUC statistic. That is, focusing on differences between the normal and cancer tissues only in the upper end of the normative range, yielded a better selection algorithm. This will not always be the case. In another set of simulations (also shown in Table 2) where informative genes had a mean 1 and standard deviation 1 in cancer tissues compared with standard normal in non-cancer tissues, the AUC statistic performed better.

These simulations assume statistical independence of genes and hence are unlikely to reflect real data. In practice one might find that subsets of informative genes are clustered statistically, and likewise subsets of uninformative genes are clustered. Intuition suggests that this would lead to higher selection probabilities than those in Table 2. We will return to simulation studies later when we consider sample size calculations.

4. Analysis of the Full Ovarian Cancer Dataset

The 1536 genes spotted on the glass arrays were ranked according to each of the discriminatory statistics defined above. Sixty-five genes were ranked in the top 100 by all 4 ranking statistics while 16 genes were selected among the top 100 by only one of the statistics (7 by ROC (0.10), 9 by pAUC (0.10), 0 by AUC and 0 by Zstat only).

The stability of their selection, quantified by $P_g(100)$, was estimated with 200 bootstrap samples. Figure 4 displays the results. The selection probabilities for the AUC and Zstat statistics are overall higher than those for the pAUC (0.10) and ROC (0.10) statistics. This presumably indicates less variability in the statistics that use more of data, namely AUC and Zstat. The selection algorithms based on them therefore are less variable and more reproducible across experiments. However, we saw earlier (Table 2) that this does not necessarily induce higher sensitivity to differential expression and in particular to the sorts of differential expression of most interest to biologists.

Another display of the resampling results, specifically for the pAUC (0.1) statistic, is shown in Figure 5. For each gene selected we calculated its ranking in each bootstrap sample. Its 80th and 90th percentile across the bootstrap samples is shown. We observe for example that gene 1483, which ranked best in the original dataset, ranked at or above 14 in 90% of the resampled datasets and at or above 8 in 80% of the resampled datasets. Gene 65, which ranked 50th in the original dataset, had ranks of 148 and 115 at its 90th and 80th bootstrap percentiles, respectively. We see that for all genes ranked in the top 24, their rankings were better than 100 in at least 90% of the bootstrap samples. Thus, we have high confidence in the good ranking of these genes, in the sense that it is unlikely to

be attributable to sampling variability. On the other hand, all of the genes that ranked worse than 63rd in the original data were at the 150th rank or worse in at least 10% of bootstrap samples, and 15/37 (41%) had 90th percentiles above 200.

Let's briefly consider the biological relevance of the highest ranking genes. The top 10 ranking clones for the pAUC(0.1) statistic are SPINT2 (2 clones), TACSTD1, HE4, Oviductal glycoprotein, Keratin 8, Argininosuccinate synthetase (ASS), 2 ESTs, and a novel gene. Of the six genes with known function, five are tumor-related: SPINT2 is expressed in colorectal cancer (Kataoka et al., 2000); TACSTD1, an adenocarcinoma-associated antigen, is currently being used in a clinical trial as a target in the treatment of gastro-intestinal carcinomas (Staib et al., 2001); HE4 is a potential ovarian cancer marker (Schummer et al., 1999), which is currently being evaluated in a serum assay (unpublished results); oviductal glycoprotein has a role in fertilization (Verhage et al., 1997) and was found to be expressed at higher levels in ovarian carcinomas (unpublished results); and Keratin 8 expression is associated with cervical cancer progression (Smedts et al., 1990). Moreover, one of the two EST-related clones is homologous to a putative integral membrane transporter protein discovered in hepatocellular carcinoma (NCBI website http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=retrieve&db=Nucleotide&list_uids=7320864&dopt=GenBank).

With five of six top-ranking genes known to be related to cancer, our biologist colleagues are highly motivated to study further the remaining 3 genes with unknown function. They

suspect that those genes are likely to be tumor-related as well. This of course remains to be seen.

5. Sample Size Calculations

Gene expression microarray experiments are expensive. Therefore in practice sample sizes tend to be small. Our simulation study and analysis of the ovarian cancer dataset nevertheless, suggested that an informative analysis, properly accounting for sampling variability, can be based on experiments with relatively small sample sizes. This initially seemed surprising to us, but in retrospect it is intuitively reasonable. Informative genes will show themselves to be so even with small sample sizes.

What advice can the statistician offer for choices of sample sizes in exploratory gene expression studies? Since the task is to select informative genes from the pool of genes studied, the criterion for choosing sample sizes should be that they be large enough to ensure that informative genes have a high chance of being selected on the basis of data from the experiment. Again, traditional notions of basing sample size calculations on hypothesis tests or on precision of estimators seem inappropriate.

Suppose that the top ranked k_0 genes will be selected for further study, $k_0 = 100$, say.

The sample size might be driven by the requirement that an informative gene, ranking in truth in the top k_1 genes ($k_1 = 30$ say), has a probability of at least β of being ranked in the top k_0 in the experiment. That is, one might choose n_D and n_c so that

$$P_g(k_0 \leq k_1) = P[\text{Rank}(g) \leq k_0 | \text{TrueRank}(g) \leq k_1] = \beta$$

where $\text{True Rank}(g)$ is the ranking of gene g according to the ranking statistic chosen, if an infinite number of tissue samples were studied, $n_D = n_c = \infty$. This probability is related to the selection probability $P_g(k)$ defined earlier. However, here instead of quantifying confidence in observed results, it now quantifies the power of the experiment to select a gene that in truth is a high ranking gene. Like $P_g(k)$ it depends on the size and contents of the pool of genes considered, the ranking statistic used and importantly on n_D and n_c .

To calculate $P_g(k_0 \leq k_1)$ we suggest that a simulation study be performed. In fact the simulation study described in section 3.3 was our first attempt at this. In that setting we calculated $P_g(100 \leq 100)$ for various sample sizes and showed that even with a total sample size of 30, $n_D = n_c = 15$, the study design had a power $\bullet = 68\%$ assuming that the pAUC (0.1) ranking statistic was used for analysis. The data generating mechanism in that simulation, however, is very simple and is not based on a theoretically justifiable model. Sample size calculations cannot be used for practical application without such justification. Unfortunately, it is extremely unlikely that one can ever stipulate a simulation model for gene expression array data that is based on adequate biological theory and knowledge of laboratory processes.

We do however have access to some real data, namely the ovarian cancer data. We based a second simulation study on these data. Specifically we resampled with replacement the

entire data vector of gene expressions for n_D^* cancer tissues and n_C^* normal tissues from the original dataset and determined $P_g(k_0 \in k_1)$. Various sample sizes, n_D^* and n_C^* were considered. That is, the distributions of observed data were regarded as the population distributions for cancers and normals and we randomly selected from those (infinite) populations in order to simulate data for the planned experiments. (In this sense bootstrapping can be considered as a simulation.) Table 3 displays $P_g(100 \in k_1)$ for various choices of sample sizes. We see that a gene that in truth ranks in the top 10 according to the pAUC (0.1) measure is almost certainly selected with data from a study involving as few as 30 tissues, if the selection criterion is that its pAUC (0.1) statistic ranks in the top 100 in the study. A gene, truly in the top 50 is likely to be selected ($\approx 91\%$) from a study using 25 cancer and 25 non-cancer tissues.

The power of $P_g(k_0 \in k_1)$ quantifies how likely a gene randomly selected from the top k_1 is likely to be ranked in the top k_0 . Table 3 also displays $P_g(k_0 \in k_1)$ which is the probability that *all* k_1 truly top-ranking genes will rank in the top k_0 when the experiment involving n_D^* and n_C^* tissues is performed. These probabilities are much lower because the criterion to be met is more stringent. In order that all top 30 genes be likely to be selected it appears that at least 100 tissues $n_D^* = n_C^* = 50$ should be studied ($P_g(100 \in 30) \approx 84\%$).

In these simulations we only considered equal numbers of cases and controls. Unequal sample sizes could be chosen. It would be interesting to see if, in general, relatively more

cases than controls are desirable and how this should in general relate to the relative variability of gene expressions in cases versus controls. Another aspect that we feel should be explored further relates to the likely over-optimism of the pilot data that we use for simulation. Efron and Tibshirani (1993, section 25.5) suggest some caution about plugging in parameters from a pilot study for power calculations and their concerns apply here too. One could add noise to the observed data in the simulations for more conservative sample size calculations.

6. Additional Design and Data Analysis Considerations

We have considered only the comparative design where assays for both the normal tissues and cancer tissues are performed, each using a common control tissue. Thus a sample of relative expression values are obtained for both the normals and the cancers, represented by $\{Y_{ig}^D, i = 1, \dots, n_D\}$, and $\{Y_{jg}^C, j = 1, \dots, n_C\}$, respectively. In this design the distribution of Y_g^D can be compared with that of Y_g^C , the latter being the appropriate reference distribution.

An alternative design frequently cited in the statistical literature (Van Der Laan and Bryan (2000)) entails using a non-cancer tissue as a control within the assay for a cancer tissue. The data at gene g from such an experiment can be represented as $\{Z_{ig}, i = 1, \dots, n_D\}$ where Z_{ig} is the expression in the cancer tissue relative to the normal

control transformed to a log scale. Typically the mean of the distribution of Z_g is compared with 0, the null value if expression at the gene g is the same on average in cancer and normal tissue.

We have argued in section 2 that the mean difference $E\{Y_g^D - Y_g^C\} = E\{Z_g\}$ is only one summary of the separation between the distributions of Y_g^D and Y_g^C , and that in many cases alternative summary measures are more relevant. Unfortunately summary measures, such as pAUC, are not identifiable from the distribution of Z_g . Indeed we believe that the two distributions for Y_g^D and Y_g^C , respectively, or at least their ROC curve should be generated by an experiment in order to adequately assess differential expression. Unfortunately they simply cannot be reconstructed from the single distribution of the composite variable Z_g . Clearly many different pairs of random variables (Y_g^D, Y_g^C) can give rise to a single composite $Z_g = Y_g^D - Y_g^C$.

In summary, for the type of application we consider in this paper, we prefer the design that yields relative expression levels for both normals and cases instead of just the composite Z_g . This design allows a full and flexible comparison of the two distributions, that for normal tissues yielding a reference distribution against which the cancer tissue distribution can be compared. Such is not achieved with the design that evaluates normals only within the assay for the cancer tissue.

7. Concluding Remarks

In this paper we have considered the identification of a subset of genes that are differentially expressed between two tissue types from a large pool of candidate genes. The same statistical problem arises in experiments involving other recently developed high throughput technologies. For example, protein mass spectrometry can be used to identify a set of proteins differentially expressed from amongst a large set of candidate proteins. Large arrays of tumor antigens are used to select a subset to which antibodies are differentially present in subjects with and without cancer. The concept of ranking genes using a statistical measure of discrimination between tissues, applies equally well to proteins in protein spectrometry and to antigens in tumor immunogenicity experiments. Thus, our methods will also be useful in these settings.

We have emphasized that investigators must carefully choose the statistical measure for ranking the genes so that it fits the purpose of the experiment. For disease screening we have argued that biomarkers must be highly specific. This could be argued for other applications too, such as in the identification of treatment targets. Statistical measures such as the $\text{pAUC}(t_0)$ or $\text{ROC}(t_0)$ statistics are appealing when specificity is important. Dudoit et al (2000b) use Zstat, the standardized difference in means, to rank genes. Efron et al (2000) also use a difference in means with a somewhat different standardization. Their rationale for using these measures over others was not discussed. One feature of those measures is that they depend on the absolute values of Y_g , whereas the empirical ROC statistics do not. This presumably infers robustness on the ROC statistics but at the expense of disallowing the magnitudes of relative expression to influence the relative

ordering of genes. Whether or not the magnitude of Y_g should influence the gene rankings over and above the separation between the probability distributions of Y_g^D and Y_g^C , is a debatable point since magnitude of expression does not translate directly into biological effect in the body. Another feature of the $\text{pAUC}(t_0)$ and $\text{ROC}(t_0)$ statistics is that they are not influenced by variability in the measurement of Y_g at the lower end of the scale, at values below the $(1-t_0)$ quantile of Y_g^C .

We have suggested the selection probability, $P_g(k)$, to quantify sampling variability and confidence in the gene ranking, and as the basis for sample size calculations. Dudoit et al (2000b) use p -values for a related purpose. However, we find the interpretation of $P_g(k)$ more compelling given the exploratory nature of the study. The purpose is not to test a null hypothesis. Efron et al (2000) consider two probabilities: a p -value, $\text{Prob}\{\text{data at gene } g \mid \text{null hypothesis of equal expression}\}$, and a Bayesian probability, $\text{Prob}\{\text{gene } g \text{ affected} \mid \text{data at gene } g\} = 1 - P\{\text{equal expression} \mid \text{data at gene } g\}$. Again, since many genes will be differentially expressed, probabilities relating to the null state of equal expression seem less compelling than ranking the extent of differential expression amongst the genes. Moreover, Efron et al (2000) use the probabilities to rank the genes, whereas we use the selection probabilities only to quantify sampling variability in the rankings.

Our selection probabilities are more closely related in this regard to the ‘single gene probabilities’ proposed by Van der Laan and Bryan (2000). The single gene probabilities

are used to quantify sampling variability in a gene clustering algorithm, and are estimated by a parametric bootstrap approach. Kerr and Churchill (2001) also assess reliability of clustering algorithms with the bootstrap. Our selection probabilities quantify sampling variability in a gene ranking algorithm, and are estimated with a non-parametric bootstrap procedure. It is likely that bootstrap or any data based estimates of $P_g(k)$ will be correlated with the data-based ordering of the gene. This correlation implies that if attention is restricted to a subset of genes that are observed to rank high say, then as a group their estimated selection probabilities will tend to be biased upwards. Efforts to reduce this bias would be worthwhile.

The initial motivation for our research was to develop a strategy for sample size calculations. The strategy we propose is based on selection probabilities for informative genes, and is implemented with bootstrap simulation studies using pilot data. A similar strategy could be used for calculating sample sizes in studies that have the determination of gene clusters as the ultimate purpose. The single gene probabilities of Van der Laan and Bryan (2000) or some related construct could take the place of the selection probabilities in that sort of application.

Although the identification of differentially expressed genes is the first objective, it is not the only objective of an exploratory gene expression study. In cancer research, it is recognized that cancer is a heterogeneous disease and that different unidentified subtypes may be characterized by unique sets of overexpressed genes. Thus, if a single gene doesn't completely discriminate cancer from non-cancer it may be possible that a small

set of genes each flagging one subtype will. Statistical methods to identify such minimal subsets are needed. Ranking of different subsets of genes might draw on ideas presented here. In addition, the identifications of clusters of genes, that is, genes that are over- or underexpressed in the same cancer tissues would be of interest. Biological insights into the pathways and pathogenesis of cancer would likely result. Some modifications of the plaid models (Lazzeroni and Owen, 2000) to include a baseline reference group of tissues (non-cancer in our case) might be useful for this purpose.

Acknowledgements

Support for this research was provided by NIH grants GM 54438 and CA 86368 and by the Fred Hutchinson Cancer Research Center Ovarian SPORC grant. We thank Lu Chen who performed the initial simulation studies, and colleagues in Seattle, particularly Nicole Urban, Ziding Feng, and David Haynor for valuable discussions.

References

Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**: 387–415.

“The Chipping Forecast,” (1999). *Nature Genetics*, **21** supplement.

Dudoit S., Fridlyand J., and Speed T.P. (2000a). Comparison of discrimination methods for the classification of tumors using gene expression data. *Technical Report*, Department of Statistics, Berkeley.

Dudoit S., Yang, Y.H., Callow M.J. and Speed T.P. (2000b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical Report*, Department of Statistics, Berkeley.

Efron B. and Tibshirani R.J. (1994). An Introduction to the Bootstrap. Chapman and Hall, CRC Press.

Efron B., Tibshirani R., Goss V., and Chu G. (2000). Microarrays and their use in a comparative experiment. *Technical Report #213*, Division of Biostatistics, Stanford University.

Hastie T., Tibshirani R., Eisen M., Brown P., Ross D., Scherf U., Weinstein J., Alizadeh A., Staudt L., and Botstein D. (2000). Gene Shaving: a new class of clustering methods for expression arrays. *Technical Report*, Department of Statistics, Stanford University.

Hintze J.L. and Nelson R.D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician* **52**(2):181–184.

Kataoka H., Itoh H., Uchino H., Hamasura N., Nabeshima K., Kono M. (2000).

Conserved expression of hepatocyte growth factor activator inhibitor type-2/placental bikunin in human colorectal carcinomas. *Cancer Letter* **148**(2):127–134.

Kerr M.K. and Churchill G.A. (submitted). Bootstrapping cluster analysis: Assessing the reliability of inclusions from microarray experiments.

Lazzeroni L. and Owen A. (2000). Plaid models for gene expression data. *Technical Report #211*, Division of Biostatistics, Stanford University.

McClish D.K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**: 190–195.

Newton M.A., Kendziorski C.M., Richmond C.S., Blattner F.R. and Tsui K.W. (2000). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computer Biology* **81**:37–52.

Pepe M.S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association* **95**: 308–311.

Pepe M.S., Etzioni R., Feng Z., Potter J., Thompson M.L., Thornquist M., Winget M., and Yasui Y. (to appear, July 2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*.

Reiser B. and Guttman I. (1986). Statistical inference for $Pr(Y < X)$: The normal case.

Technometrics **28**: 253-257.

Schummer M., Ng W.V., Bumgarner R.E., Nelson P.S., Schummer B., Bednarski D.W., Hassell L., Baldwin R.L., Karlan B.Y., Hood L. (1999). Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Genetics* **238**(2):375–385.

Smedts F., Ramaekers F., Robben H., Pruszcynski M., van Mujien G., Lave B. Leigh I., Vooijs P. (1990). Changing patterns of keratin expression during progression of cervical epithelial neoplasia. *American Journal of Pathology* **136**(3):657–668.

Staib L., Birebent B., Somasundaram R., Purev E., Braumuller H., Leeser C., Kuttner N., Li W., Zhu D., Diao J., Wunner W., Speicher D., Beger H.G., Song H., Herlyn D. (2001). Immunogenicity of recombinant GA733-2E antigen (CO17-1A, EGP, KS1-4 Ep-CAM) in gastro-intestinal carcinoma patients. *International Journal of Cancer* **92**(1):79–87.

Tibshirani R., Hastie T., Eisen M., Ross D., Botstein D., and Brown P. (2000). Clustering methods for the analysis of DNA microarray data, *Technical Report*, Department of Statistics, Stanford University.

Van der Laan M. and Bryan J. (2000). Gene expression analysis with the parametric bootstrap. *Technical Report*, Division of Biostatistics, Berkeley.

Verhage H.G., Fazleabas A.T., Mayrogiannis P.A., O'Day-Bowman M.B., Schmidt A.,
Arias E.B., Jaffe R.C. (1997). Characteristics of an oviductal glycoprotein and its
potential role in fertility. *Journal of Reproduction and Fertility* (Supplement) **51**:217–
226.

Figure Legends

Figure 1. Hypothetical distributions for gene expression data showing different sorts of separations between cancer tissue and normal tissue.

Figure 2. Receiver operating characteristic curves corresponding to pairs of distributions shown in Figure 1.

Figure 3(a). Frequency distributions and (b)ROC curves corresponding to gene 5 and 97 in the ovarian cancer data set.

Figure 4. (a) Violin plots (Hintze and Nelson, 1998) of selection probabilities for the top 100 ranked genes in the ovarian cancer datasets using 4 different ranking statistics. Probability estimates are based on 200 bootstrap samples. The median is indicated by a short horizontal line, the first to third interquartile range by the narrow shaded box, and a vertical line extends to the upper and lower adjacent values. The surrounding violin shell consists of mirrored local kernel density estimates of the distribution. The y-axis is labeled at the minimum, median, and maximum values..

(b) A comparison of the selection probabilities for the AUC statistic and the pAUC (0.1) statistic.

Figure 5. Gene rank percentiles (90th and 80th) in the bootstrap distribution for the ovarian cancer data set. Shown are results for the top ranked 100 genes. 200 bootstrap samples were drawn with the sampling unit being tissue.

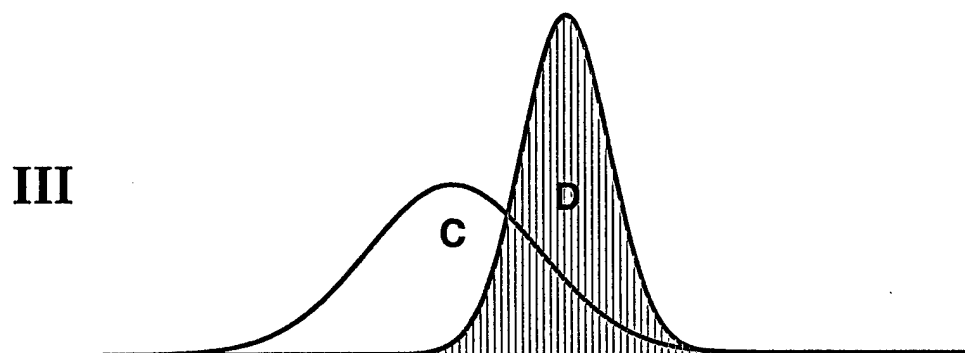
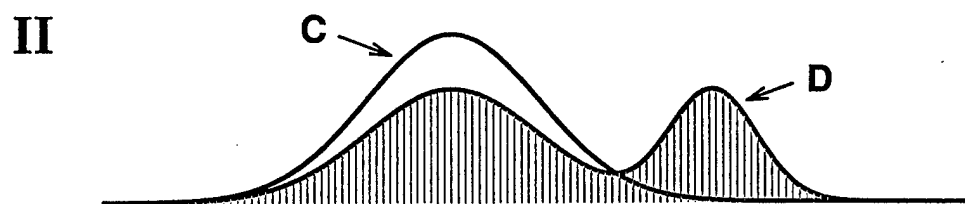
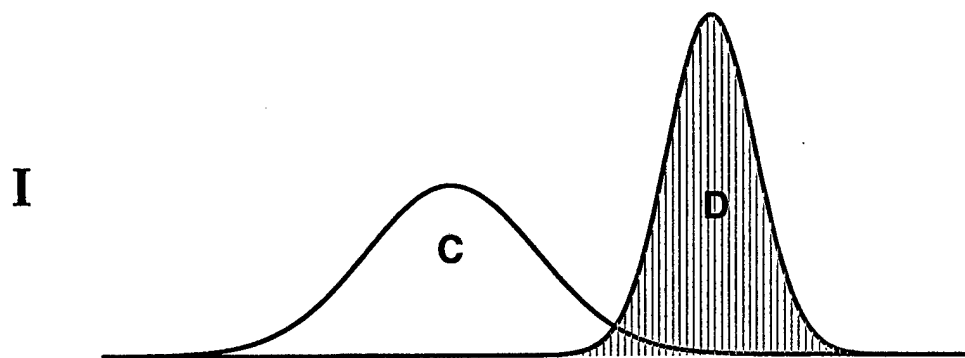


Figure 1

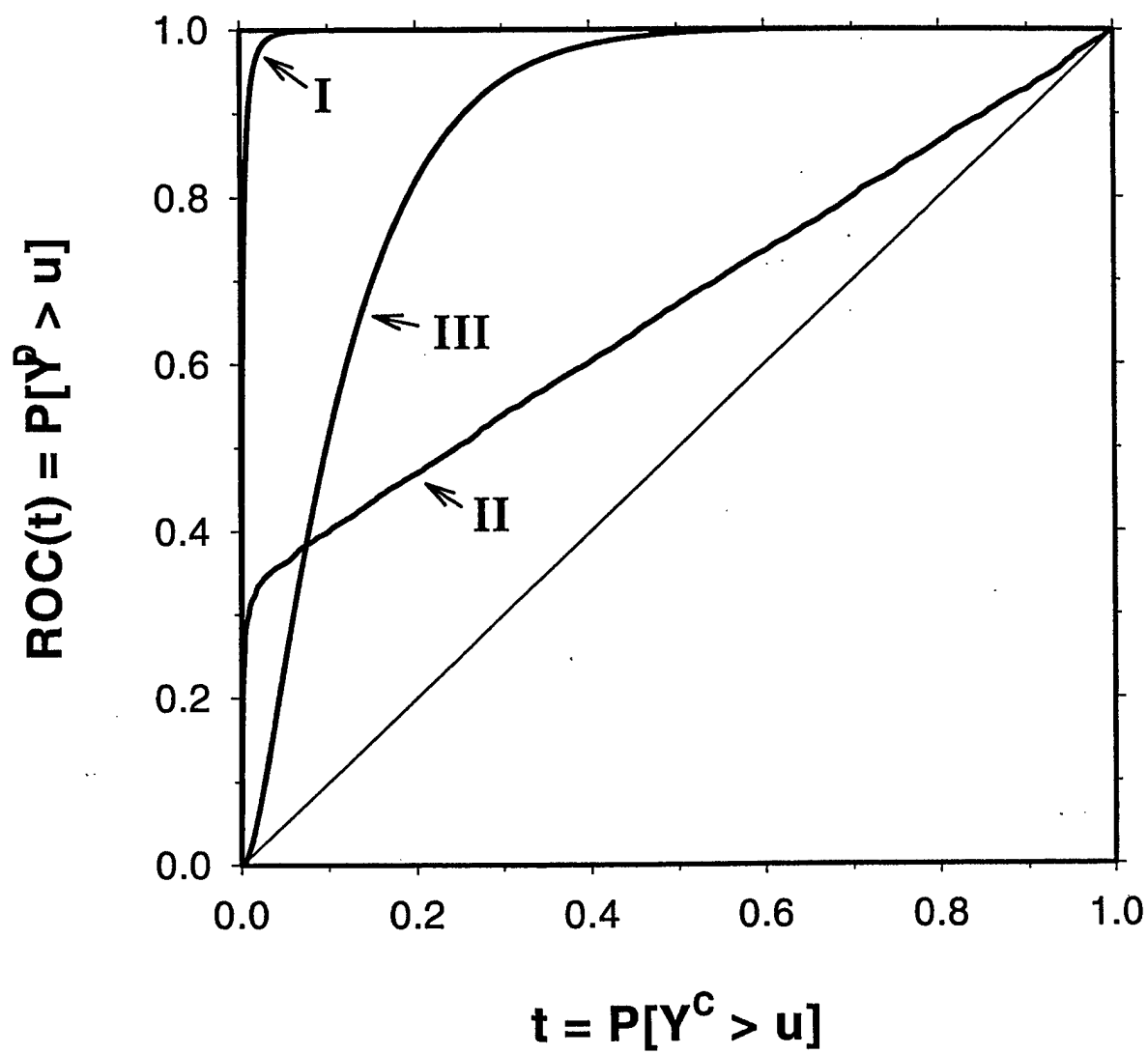
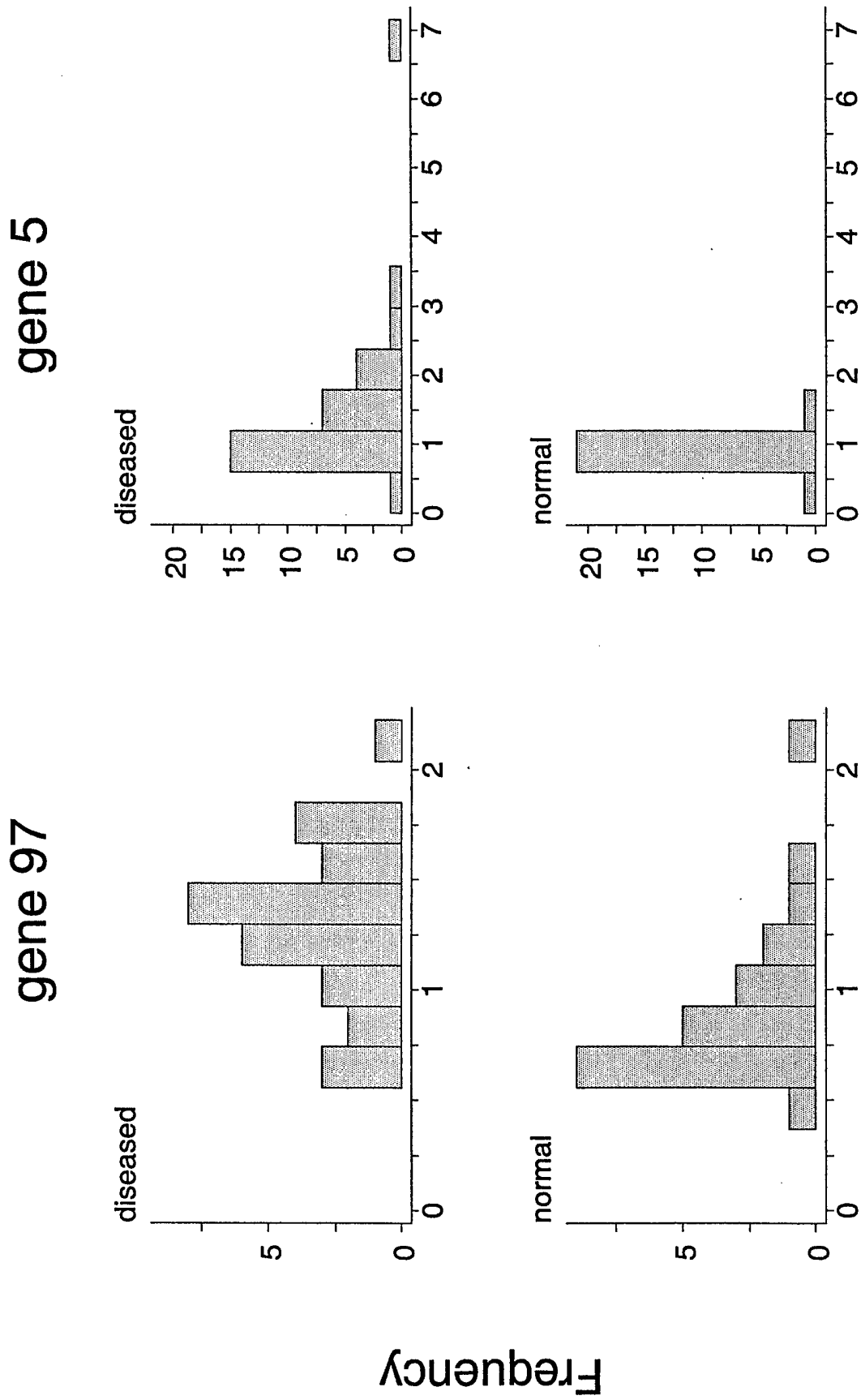


Figure 2



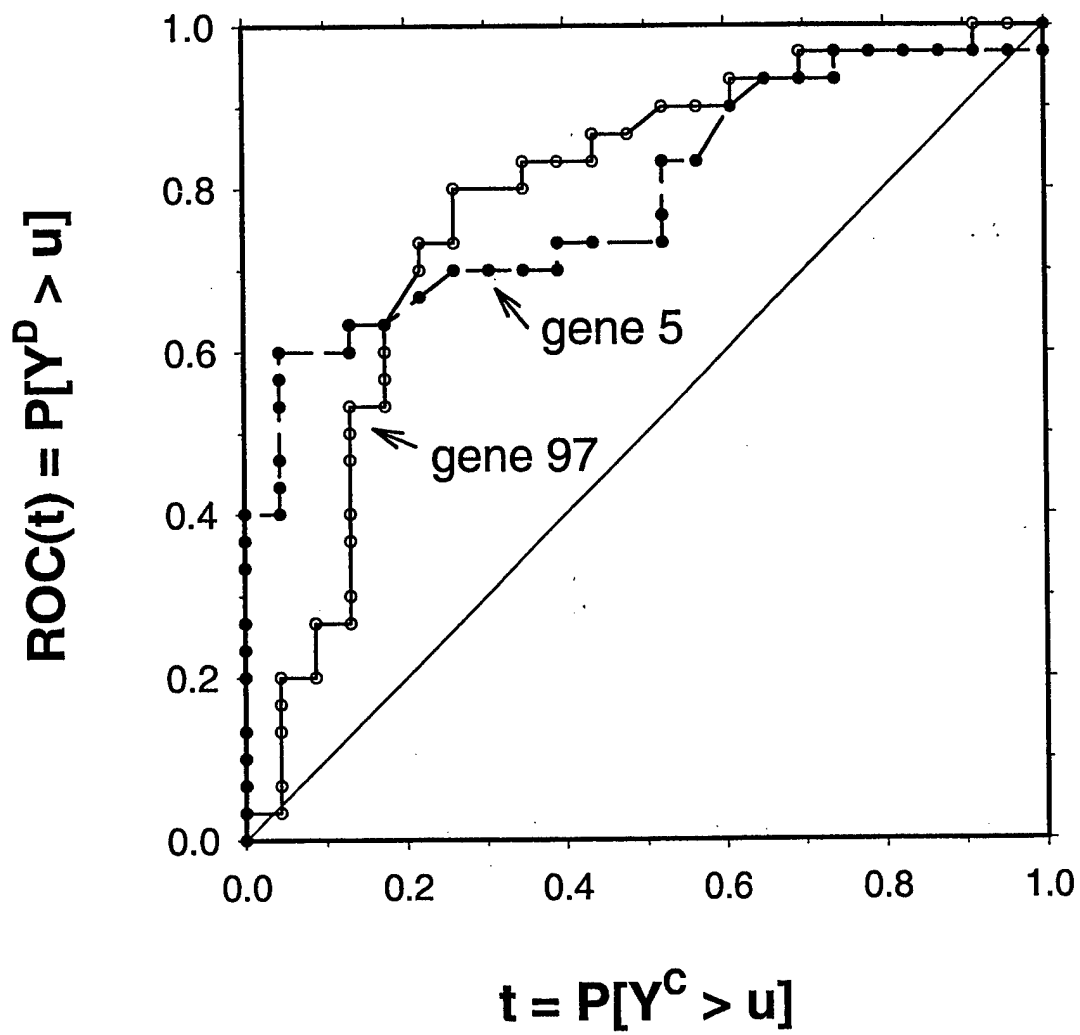
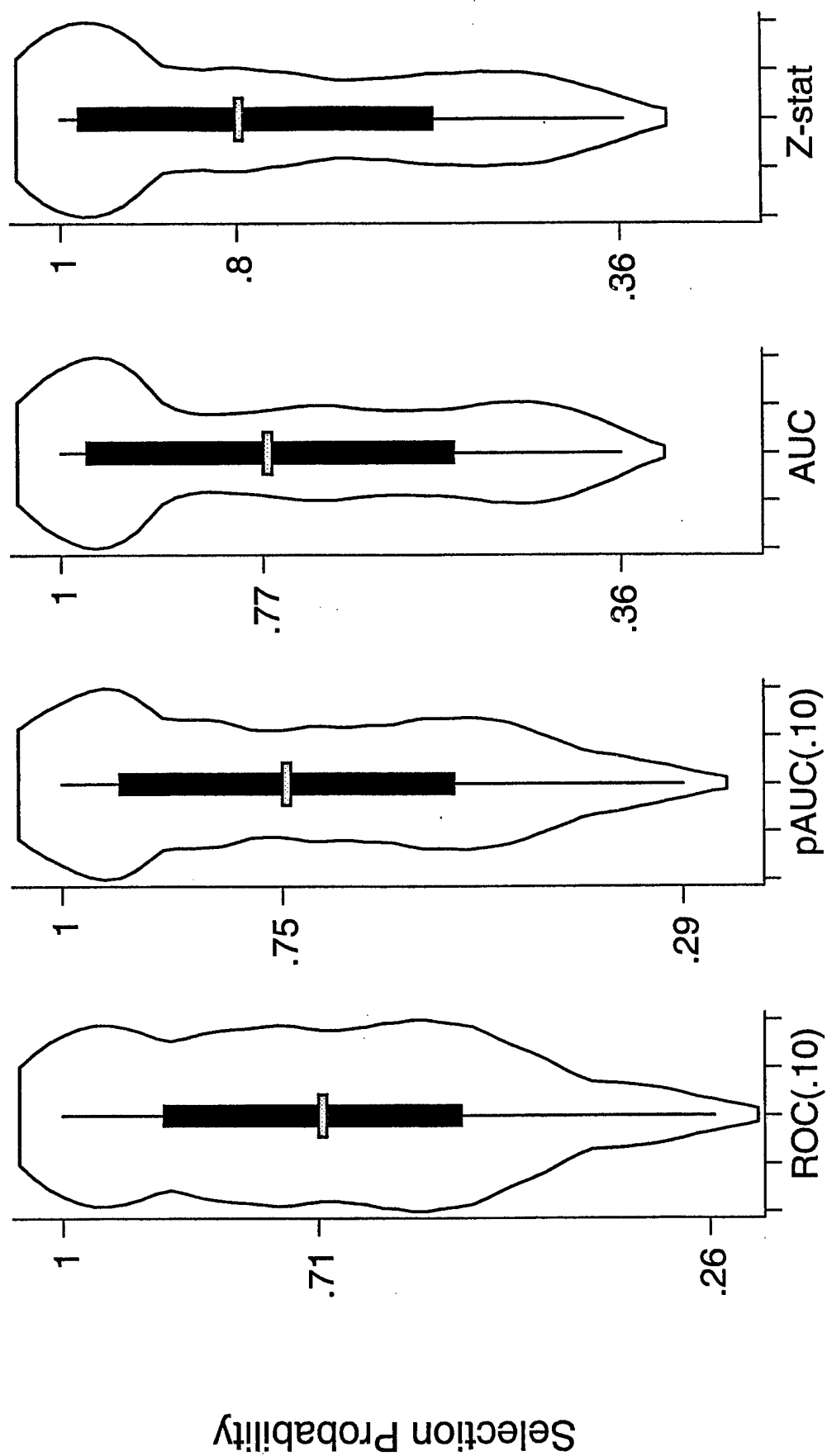


Figure 3b

Figure 4a



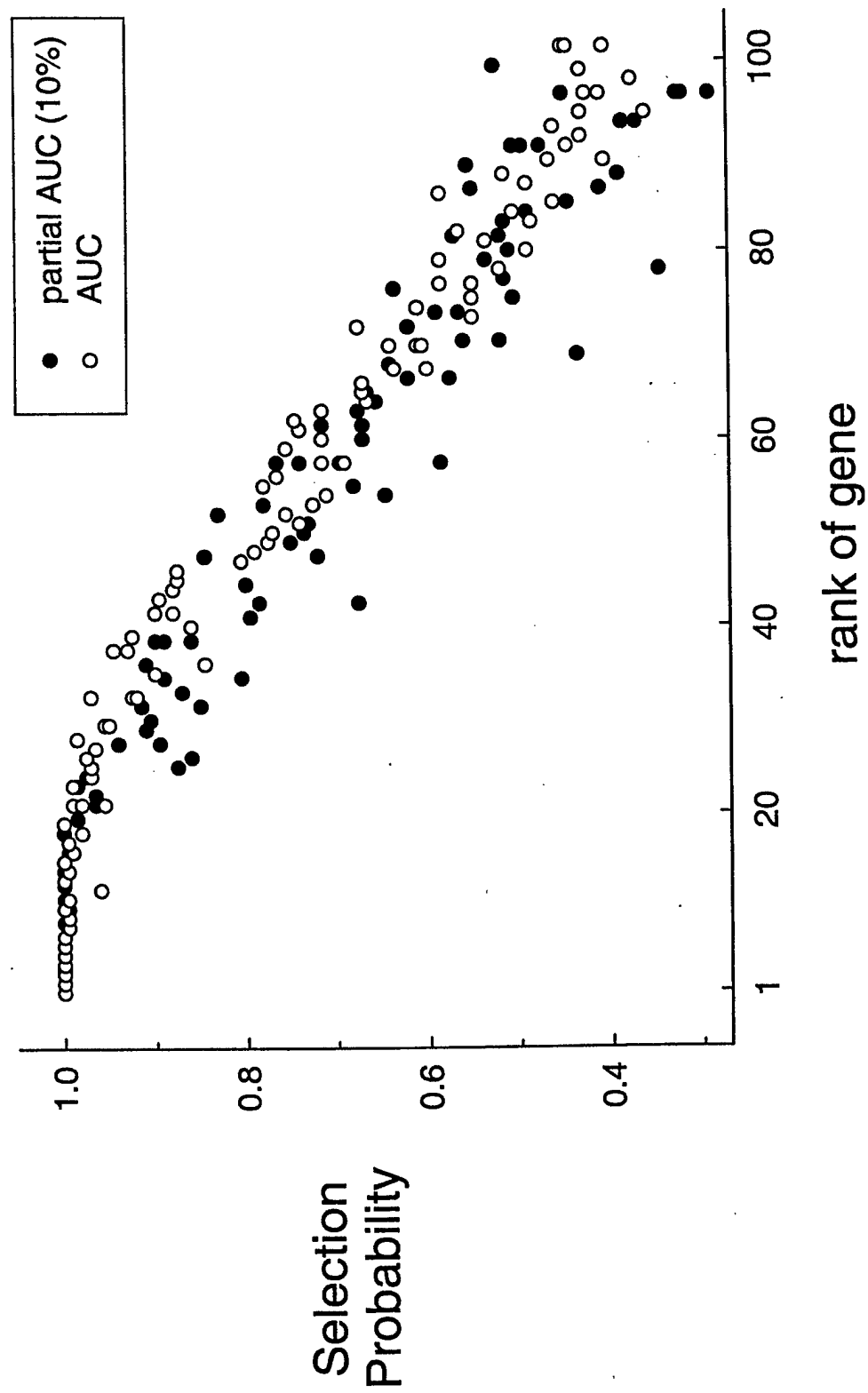


Figure 4b

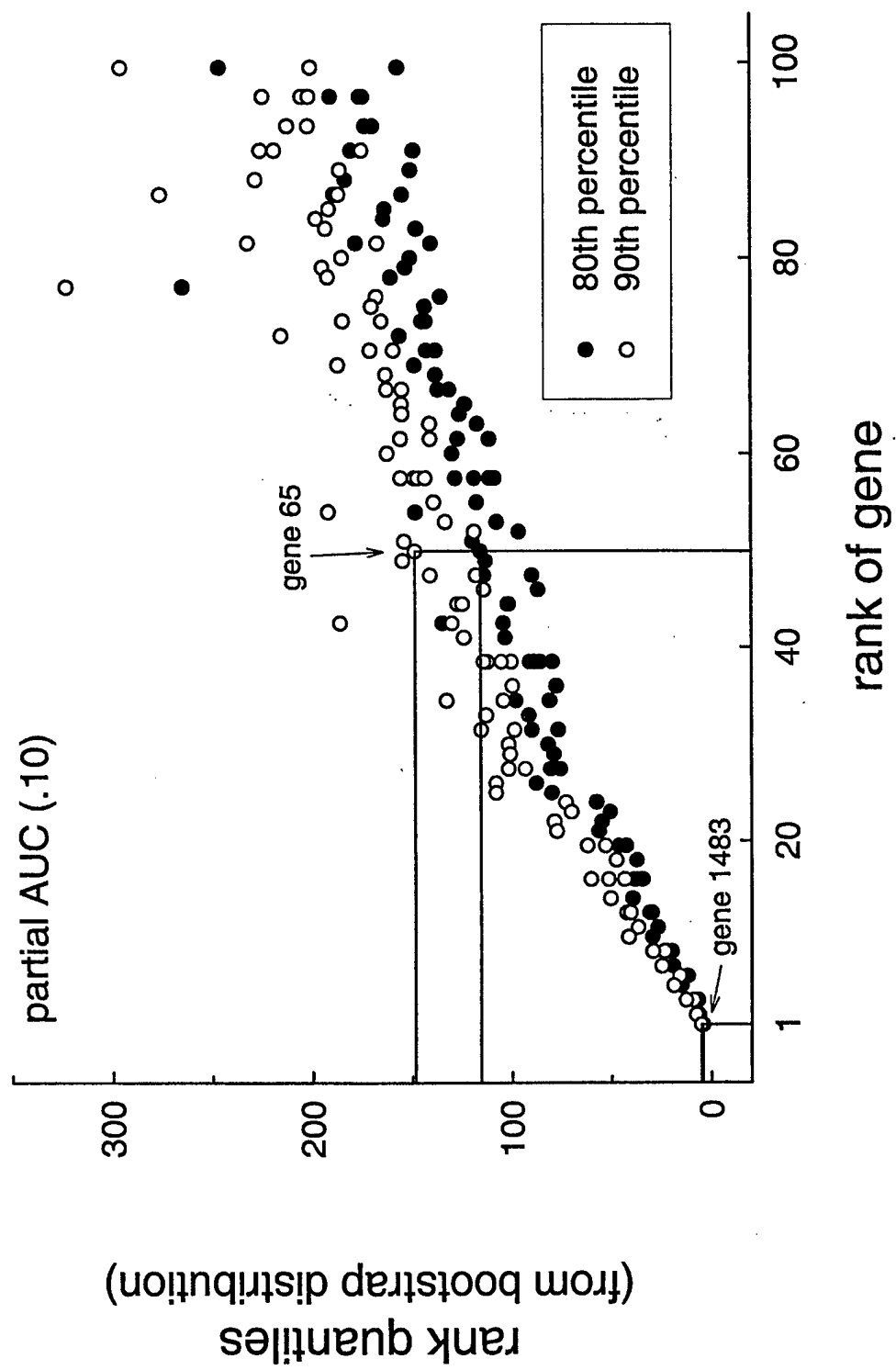


Table 1.
Gene number, selection probability $P_g(10)$, and value of the discriminatory measure for the top 10 ranking genes among the first 100 genes in the ovarian cancer dataset.

rank	ROC(.10)				pAUC(.10)				AUC				Z-stat			
	$P_g(10)$		statistic		gene		$P_g(10)$		statistic		rank		gene		$P_g(10)$	
	gene		rank		gene		rank		gene		rank		gene		rank	statistic
1	93	1.00	0.900	1	93	1.00	0.090	1	93	1.00	0.971	1	93	1.00	9.493	9.493
2	76	0.81	0.767	2	65	0.99	0.059	2	42	1.00	0.870	2	65	1.00	8.445	8.445
3	65	0.98	0.733	3	5	0.94	0.051	3	76	1.00	0.864	3	42	0.94	7.425	7.425
4	42	0.92	0.667	4	23	0.83	0.044	4	65	0.98	0.854	4	97	0.74	6.466	6.466
5	5	0.89	0.600	5	42	0.60	0.041	5	16	0.82	0.804	5	39	0.66	6.136	6.136
6.5	16	0.71	0.533	6	51	0.68	0.040	6	5	0.74	0.789	6	23	0.61	5.686	5.686
6.5	39	0.61	0.533	7	52	0.63	0.040	7	52	0.74	0.784	7	35	0.55	5.652	5.652
8	35	0.58	0.500	8	35	0.38	0.033	8	97	0.71	0.780	8	76	0.50	5.085	5.085
9.5	23	0.54	0.467	9	73	0.38	0.032	9	39	0.52	0.752	9	63	0.32	4.922	4.922
9.5	52	0.43	0.467	10	76	0.48	0.032	10	75	0.43	0.736	10	5	0.47	4.899	4.899

Table 2

Results of a simulation study with $N=2000$ genes of which 100 are informative about disease status. Shown are $P[\text{Rank}(g) > 100]$ for informative genes. In all simulations Y_g has a standard normal distribution among controls and for non-informative genes among cases. The distribution of Y_g^D for informative genes is $N(1,2)$ in setting A and $N(1,1)$ in setting B.

Statistic	A			B		
	n=#cases=#controls			n=#cases=#controls		
	15	25	35	15	25	35
ROC(.10)	.69	.82	.89	.57	.69	.77
pAUC(.10)	.68	.83	.92	.50	.62	.72
ROC(.20)	.59	.75	.83	.65	.76	.84
pAUC(.20)	.68	.83	.91	.58	.71	.81
AUC	.42	.56	.66	.68	.84	.92
T statistic	.43	.58	.68	.69	.85	.92

Table 3

Study power $P_g \{100| \in k_1\}$ as a function of sample size using the ovarian cancer data as a simulation model. Also shown is the power for the more stringent criterion $P_g \{100| \cup k_1\}$.

True Ranking (k_1)	$P_g \{100 \in k_1\}$				
	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50
(n_D, n_C)					
(15, 15)	.997	.982	.934	.893	.850
(25, 25)	1.000	.996	.973	.949	.914
(50, 50)	1.000	1.000	.994	.987	.968
(100, 100)	1.000	1.000	.999	.998	.990
	$P_g \{100 \cup k_1\}$				
(15, 15)	.960	.654	.120	.016	.000
(25, 25)	1.000	.928	.486	.202	.024
(50, 50)	1.000	1.000	.836	.638	.206
(100, 100)	1.000	1.000	.984	.928	.608

A Parametric Empirical Bayes Method for Cancer Screening using Longitudinal Observations of a Biomarker

BY MARTIN W. MCINTOSH

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue MP-900, Seattle, Washington 98109
e-mail: mmcintos@fhcrc.org

NICOLE URBAN

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue MP-900, Seattle, Washington 98109
e-mail: nurban@fhcrc.org

SUMMARY

A revolution in molecular technology is leading to the discovery of many biomarkers of disease. Monitoring these biomarkers in a population may lead to earlier disease detection, and may prevent death from diseases like cancer that are more curable if found early. For markers whose concentration is associated with disease progression the earliest detection is achieved by monitoring the marker with an algorithm able to detect very small changes. One strategy is to monitor the biomarkers using a longitudinal algorithm that incorporates a subject's screening history into screening decisions. Longitudinal algorithms that have been proposed thus far rely on modeling the behavior of a biomarker from the moment of disease onset until it's clinical presentation. Because the data needed to observe the early pre clinical behavior of the biomarker may take years to accumulate, those algorithms are not appropriate for timely development using new biomarker discoveries. This manuscript presents a computationally simple longitudinal screening algorithm that can be implemented with data that is obtainable in a short period of time. For biomarkers meeting only a few modest assumptions our algorithm uniformly improves the sensitivity compared to simpler screening algorithms but maintains the same specificity. It is unclear what performance advantage more complex methods may have compared to our method, especially when there is doubt about the correct model for describing the behavior of the biomarker early in the disease process. Our method was specifically developed for use in screening for cancer with a new biomarker, but it is appropriate whenever the pre-clinical behavior of the disease and/or biomarker is uncertain.

Some key words: ovarian cancer, change-point model, PEB algorithm, CA 125

1. INTRODUCTION

A revolution in molecular technology is leading the discovery of a large number of genes that are highly expressed or amplified in cancer. If a gene's protein product, a tumor marker, can be found elevated in the blood of diseased subjects then that tumor marker may be useful for cancer screening. Thus, with numerous gene discoveries come numerous candidate tumor markers. Translating a tumor marker discovery from the laboratory to use in a population requires stating a screening algorithm. The most commonly used algorithm, the single threshold (ST) rule, classifies a subject positive (likely to have disease) when his or her marker concentration exceeds a common population-wide threshold. A better approach may be to use a longitudinal screening algorithm that makes use of a subject's accumulated screening history in order to achieve even earlier detection. This manuscript describes a formal procedure for generating longitudinal algorithms appropriate for new marker discoveries.

Tumor markers are useful for early detection if after cancer onset, and prior to clinical presentation, their concentration in serum or other fluids deviates from their normal levels. Previously proposed longitudinal algorithms work by monitoring their concentration over time and give a positive screen to subjects whose trajectories appear more similar to that suspected in pre clinical disease than normal. To make this assessment, however, those algorithms must model the pre clinical marker trajectory from the moment of disease onset through its clinical presentation. For example, Slate and Cronin (1997), Slate and Clark (1999), and Morrell *et al.* (1995) screen for prostate cancer with the PSA tumor marker and Skates and Pauler (2001) screen for ovarian cancer with the CA 125 tumor marker. If the early pre-clinical behavior of the tumor marker is modeled correctly, these algorithms yield optimal screening decisions. Unfortunately, obtaining sufficient data to estimate the pre clinical marker trajectory is impractical for many diseases, especially those with low prevalence and rapid progression.

Because longitudinal algorithms are intended to detect cancer *earlier* than the ST rule they must do so by giving a positive screen when marker concentrations are lower than the ST rules' threshold. Thus, the behavior after marker trajectory exceeds the ST rule is not relevant for earlier detection, and so any approach that models marker trajectory must be particularly accurate in its representation of the very early pre clinical period. Of course, observing the early trajectory requires having access to a sufficient number of serial specimens collected during this time. Unfortunately, such observations are extremely rare because by definition a screened population has low disease prevalence (screened populations are asymptomatic). Even sera stored as part of a large cohort study may not be sufficient. For example, a recent ovarian cancer screening trial (Jacobs *et al.*, 1999) collected serum specimens at each of three annual screens from 10,958 women. Specimens were also collected every three months for up to a year from women whose CA 125 concentration was found above its ST rule of 30 U/ml. Only six women were found with ovarian cancer

during the screening period, with ten more found during the seven year followup. Because the pre-clinical period of ovarian cancer is thought to be between 10 months (Bast *et al.*, 1985) and 18 months (Zurawski *et al.*, 1988), few of those ten cases are likely to contain any serial samples collected during the pre-clinical period. Although serial samples may have been captured from the other six cases, the screening interval may have been too short to observe the trajectory of this rapidly developing disease. Moreover, although the six screen detected cases contribute serial specimens following initial elevation, that observed trajectory is not informative about the earlier behavior when it was below 30 U/ml, the part most relevant if earlier detection is to be achieved. Finally, the information contained in any stored specimens, even if they are abundant enough, may not be accessible for new markers because specimen needs (i.e., plasma or urine instead of sera) and/or specimen-processing requirements of new discoveries cannot be anticipated.

Because the early pre clinical period may be unobservable, algorithms that do not rely on knowing it may offer a valuable alternative. One simple screening algorithm that uses screening history but does not depend on the assumed marker trajectory has been proposed by Urban *et al.* (1997). Using a microsimulation model, they found that detection rates for ovarian cancer were improved by giving a positive screen when CA125 has either doubled since the most recent screen or has reached a threshold. Because no formal (statistical) framework was used no general statements about its efficiency can be made, and their conclusions cannot be generalized to other markers or diseases. Indeed under some circumstances the ST rule may outperform a doubling rule (see Section 4).

Our method uses screening history to decide if a general, rather than a specific (i.e., exponential or linear), change in marker behavior has occurred, and does not require stating a specific marker growth model. Our method is motivated by the need for an algorithm that is robust to a variety of growth behaviors, and to have specimen demands that are practical. Indeed, our approach uses specimens that are quite abundant because its attention is focused on the healthy subjects.

Our screening rule uses Parametric Empirical Bayes (PEB) theory to generalize the ST rule to be conditional on screening history. Each subject's screening history is used to set an individually tailored threshold that achieves an arbitrary false positive rate. Controlling the false positive rate is particularly appropriate for cancer screening because costs (social and economic) are dominated by the rate of positive tests which, because disease prevalence is low in screening, is dominated by the error rate in the healthy subjects (false positive rate). Thus, an overarching constraint of any screening program is to limit the error rate in the healthy subjects, and any screening algorithm, whatever its detection rate or approach, must accommodate it.

With the PEB screening rule we find that a great majority of all subjects can have their cancer detected with marker concentration at levels far lower than with the ST rule, and so cancer is detected earlier.

Importantly, this earlier detection is achieved while maintaining the same population false positive rate and so the overall cost of a screening program is controlled. The PEB rule uniformly dominates the ST rule, and can be used with any marker having some mathematical transformation that gives the healthy subjects a hierarchical normal marker distribution. These assumptions appear widely appropriate for many tumor markers (Crump *et al.*, 2000; Pauler *et al.*, 2001). No specific restrictions are made on the behavior of the markers in the cases other than that their concentrations deviate (elevate or fall) from normal following cancer onset. Specimen requirements are modest, as serial measurements on cases are not needed.

Section 2 outlines the notation and statistical model assumed throughout the text. Section 3 uses the statistical model to derive the ST screening rule most commonly used in cancer screening, and then generalizes it to person specific thresholds using PEB theory. All screening rules in Section 3 are defined to have comparable performance in healthy subjects, and the performance of the algorithm in diseased subjects is discussed by Section 4. Section 5 presents an example using CA 125 to screen for ovarian cancer. Section 6 provides a discussion of the robustness of our algorithm, and presents technical details that may be skipped. A summary is given by Section 7

2. BACKGROUND AND NOTATION

We use Y_j to represent the marker concentration at a person's j^{th} screen, and D_j to represent the disease status at that time; $D_j = 0$ when disease free, and $D_j = 1$ otherwise. We assume cancer does not regress, so $D_j = 1$ implies $D_{j'} = 1$ for all $j' > j$. We use lower case letters to denote observed quantities, and so the screening history from a subject having completed n screens is denoted y_1, \dots, y_n . Section 2.1 represents the statistical model used to represent the behavior of the marker when without cancer (when $D_j = 0$), and Section 2.2 represents the behavior of the marker after cancer onset (when $D_j = 1$). When without cancer our marker behaves in a manner similar to that used by Skates and Pauler (2001), but whereas their method is for only a specific pre-clinical marker behavior, ours requires only that the marker deviates (elevates or falls) following cancer onset.

2.1. Marker behavior in subjects without cancer

We represent the marker concentration at the j^{th} screen for a healthy subject with the following hierarchical model.

$$Y_j | \mu, \sigma \sim N(\mu, \sigma^2) \quad (1)$$

$$\mu | \bar{\mu}, \tau \sim N(\bar{\mu}, \tau^2) \quad (2)$$

which implies that the marginal distribution of the marker in the population of healthy subjects is

$$Y_j | \bar{\mu}, \tau, \sigma \sim N(\bar{\mu}, \tau^2 + \sigma^2) \quad (3)$$

Note that the distributions above are implicitly conditioning on $D_j = 0$.

The individuals in a population are given person specific normal marker levels denoted by μ (we suppress person-level subscripts for μ and Y_j for clarity), but within subject variances, σ^2 , are assumed equal. $\bar{\mu}$ and τ denote the mean and standard deviation of μ in the population, respectively.

Screening rules are invariant to monotonic transformation so generality is not lost by assuming distribution (3) for any continuous marker because then some transformation to normality is assured. Although our hierarchical model is similar to that used by other more complex methods (see Slate and Cronin, 1997; Pauler *et al.*, 2001; Skates and Singer, 1991), the existence of suitable transformation for those methods is not assured because they require the transformation to be compatible with their specific marker growth model.

We can further reduce the total number of parameters in the model by re centering and rescaling (both monotonic transformations) so that the marginal distribution (3) has zero mean and unit variance. This *reduced model*, is given by

$$Y_j | \mu \sim N(\mu, 1 - B_1) \quad (4)$$

$$\mu | B_1 \sim N(0, B_1) \quad (5)$$

where $B_1 = \frac{\tau^2}{\sigma^2 + \tau^2}$. The marginal distribution (3) becomes

$$Y_j \sim N(0, 1) \quad (6)$$

The reduced model (4)–(6) has only a single parameter, B_1 , which represents the fraction of total variability accounted for the the between woman variance. The reason for the subscript becomes apparent below. Because it has fewer parameters we will prefer it to the full model (1)–(3) when deriving our screening rules in Section 3.

The model above and the screening rules below can be extended to accommodate fixed and time varying covariates (e.g., 'race' for the CA 125 ovarian cancer marker and 'age' for the prostate cancer marker PSA), and serial correlation. Time varying covariates and serial correlation must be accommodated if their effects are substantial in order to avoid bias. Including fixed covariate improve the overall precision of the screening rule. Accommodate fixed covariates by replacing $\bar{\mu}$ with a linear predictor $\bar{\mu} + X\beta$ (see Morris, 1983b; Pauler *et al.*, 2001). Accommodate time varying covariates by modeling the residuals, or deviations, from the population expected value. Specifically, if X_t represents the covariate value at time t , and $\bar{\mu} + X_{it}\beta$ predicts the mean

marker concentration for all subjects with covariate X_t (i.e., age), then the residual $R_{it} = Y_{it} - X_{it}\beta$ follows the hierarchical model (1)–(3). Accommodate serial correlation by inflating (deflating) the within person variance in (1) or (4) based on the time-dependence model used. For the remainder of this manuscript we assume that these effects have been accommodated so that the hierarchical model (4)–(6) is an adequate representation of the marker trajectory in healthy subjects.

2.2. Marker behavior with the onset of cancer

After cancer onset a marker may find its mean level deviate from μ by an amount δ_t , where $t \geq 0$ represents the time since cancer onset. We assume $0 < \delta_t < \delta_{t'}$ when $t < t'$ because most typically markers grow after cancer onset (see Baron *et al.*, 1999, for a rare exception. Our method is also appropriate for markers that descend after onset). Thus, when a subject has cancer the expected value of their marker is given by $\mu + \delta_t$.

3. SCREENING RULES

The ST rule defines a single value and gives a positive screen to any subject found to have a marker that exceeds it. The threshold is chosen to achieve a specified population-wide false positive rate (FPR), denoted f_0 . Formally, the ST rule gives a positive screen whenever $Y_j > c(f_0)$ where $c(f_0)$ satisfies $f_0 = P(Y_j > c(f_0) | D_j = 0)$. In contrast, a longitudinal algorithm determines a threshold that depends on the subject's screening history y_1, \dots, y_n . Formally, a positive screen is given to screen $n + 1$ if $Y_{n+1} > c(y_1, \dots, y_n; f_0)$, where the function $c(y_1, \dots, y_n; f_0)$ is determined by a particular choice of algorithm. This section shows how a longitudinal algorithm can be generated by naturally generalizing the ST rule while maintaining the overall false positive rate at $f_0 = P(Y_{n+1} > c(y_1, \dots, y_n; f_0) | D_n = 0, y_1 \dots y_n)$.

We generate screening rules using the logic of hypothesis tests (see McIntosh and Pepe, 2001, for the equivalence of hypothesis testing and screening test generation). All screening rules below use the reduced model (4)–(3) and assume the model parameters $\bar{\mu}, \tau, \sigma$ and the transformation to marginal normality are known, or estimated with high precision. The actual number of specimens required to estimate the transformation and model parameters depends on the intended specificity of the screening algorithm; high specificity implies that the tail of the distribution must be accurate, and will increase data demands. However, because only serial samples on healthy subjects are needed, sufficient data can be accumulated in a short period of time. If needed, the models can be extended to allow for uncertainty in them (perhaps using a fully Bayesian framework, or with a mixed linear model and using Best Linear Unbiased Predictors (BLUPs)).

3.1. Single Threshold (ST) screening rule

The ST rule is generated by the marginal distribution (6) by carrying out the hypothesis test $H_0 : E(Y_j) = \bar{\mu}$ versus $H_a : E(Y_j) > \bar{\mu}$ with type I error rate equal to f_0 . The ST rule gives a positive screen whenever $Y_j > z_{1-f_0}$ where z_{1-f_0} is the $100 \times (1 - f_0)$ percentile of the normal distribution.

3.2. An individually tailored screening rule

If μ were known perfectly the ST rule could be tailored to the individual by performing the hypothesis test $H_0 : E(Y_j|\mu) = \mu$, versus $H_a : E(Y_j) > \mu$. Because the between subject source of variation is eliminated (equals B_1 in the reduced model) this *limiting rule* screens positive whenever

$$Y_j \geq \mu + z_{1-f_0} \sqrt{1 - B_1} \quad (7)$$

Intuitively we see that this limiting rule outperforms the ST rule because it detects deviations from normal that are $\sqrt{1 - B_1} \times 100\%$ the size but with the same false positive rate. We show this formally in Section 4.

3.3. A naïve sequential rule

Because μ is not known we must find a compromise rule than uses information about μ obtained over time. A subject undergoing an $n + 1^{th}$ screening event has a history of length n , and we summarize that history with $\bar{y}_n = \sum_{j=1}^n y_j/n$. A simple approach to control for this history replaces μ in (7) with \bar{y}_n , but adjusts for the uncertainty in \bar{y}_n as an estimate of μ in order to maintain $FPR=f_0$. In the reduced model this *naïve sequential* (NS) rule gives a positive screen whenever

$$Y_{n+1} > \bar{y}_n + z_{1-f_0} \sqrt{(1 - B_1)(1 + 1/n)} \quad (8)$$

Note that the NS rule has the same form as the ST and limiting rule but with the factor $\sqrt{(1 - B_1)(1 + 1/n)}$. The ST rule (when $n = 0$) and the limiting rule (when $n = \infty$) are special cases of the NS rule. Intuitively we see that the NS rule can do better or worse than the ST rule depending on whether $\sqrt{(1 - B_1)(1 + 1/n)}$ is less than or greater than unity, which is not assured when n , or B_1 are small. For example, if $B_1 = 1/4$ then not until $n = 5$ will the naïve rule meet or exceed the performance of the ST rule.

3.4. PEB Screening Rule

The NS rule replaces μ in (7) by its usual unbiased estimate, \bar{y}_n . Here we propose replacing μ by its Parametric Empirical Bayes (PEB) estimate. PEB refers to class of statistical procedures for estimating a group of individual means when drawn from a common population (for example see Morris, 1983b; Casella, 1985;

Efron and Morris, 1997, for technical and non-technical overviews of PEB methods). The PEB estimator, denoted $\hat{\mu}_n$, is a function of the observed sample mean \bar{y}_n and the population parameters. The estimator is expressed as

$$\hat{\mu}_n = \bar{\mu}(1 - B_n) + \bar{y}_n B_n \quad (9)$$

where the shrinkage factor B_n equals

$$B_n = \frac{\tau^2}{\sigma^2/n + \tau^2}$$

Setting $\bar{\mu} = 0$ in (9) gives the PEB estimate for the reduced model. Although the expression (9) has the form of the familiar Bayes estimator we call it an Empirical Bayes estimator because the “prior” parameters are estimated empirically, and so the properties we attribute to it are frequentist in nature.

Because the shrinkage factor, B_n , lies between zero (when $n = 0$) and one (when n is large) the PEB estimator is a compromise estimator that falls between the population mean and the individual’s sample average. Because the shrinkage factor increases with n the PEB estimate comes into closer agreement with the individual average when history is plentiful. With small n the PEB estimator is closer to the population average. Thus, the PEB estimator allows the individual history to carry a greater voice when history is substantial, but anticipates regression for subjects having little screening history.

The PEB estimator has two properties that make it useful for screening.

- (i) it is an unbiased estimate of a future observation just like \bar{y}_n ,

$$E(Y_{n+1} | \bar{y}_n) = \hat{\mu}_n$$

- (ii) but it has smaller variance

$$\text{Var}(\hat{\mu}_n | \bar{y}_n) = \text{Var}(\bar{Y}_n) B_n = \frac{\sigma^2}{n} B_n \leq \frac{\sigma^2}{n} = \text{Var}(\bar{Y}_n)$$

In the reduced model we express this variance as

$$\text{Var}(\hat{\mu}_n | \bar{y}_n) = \frac{1}{n} (1 - B_1) B_n$$

The expectations in (i) and (ii) are *conditioned* on $\bar{Y}_n = \bar{y}_n$ and the population parameters. See Morris (1983a) for derivations, but note that the parameterization of B_1 and B_n used here differ slightly. When no covariates are used these properties may be derived using normal regression formulae as well. For example, derive property (i) by noting $\text{Cov}(Y_{n+1}, \bar{Y}_n) = \tau^2$ and $\text{Var}(Y_n) = \tau^2 + \sigma^2/n$, and so regressing Y_{n+1} on \bar{Y}_n gives the coefficient $\frac{\tau^2}{\tau^2 + \sigma^2/n} = B_n$. The advantage of the PEB viewpoint over the regression viewpoint is that generalization to covariates is straight forward.

We generalize the ST rule with the PEB estimator using the hypothesis test: $H_0 : E(Y_{n+1}|\hat{\mu}_n) = \hat{\mu}_n$ versus $H_a : E(Y_{n+1} | \hat{\mu}_n) \geq \hat{\mu}_n$, and reject when Y_{n+1} exceeds its *conditional* expectation by too much. The distribution of $Y_{n+1} - \hat{\mu}_n$ under the null is normal, with variance given by

$$\text{Var}(Y_{n+1} - \hat{\mu}_n | \bar{y}_n, D_{n+1} = 0) = \text{Var}(Y_{n+1}|\bar{y}_n) + \text{Var}(\hat{\mu} | \bar{y}_n) \quad (10)$$

$$\begin{aligned} &= (1 - B_1) + \frac{1}{n}(1 - B_1)B_n \\ &= 1 - B_1B_n \end{aligned} \quad (11)$$

Expression (10) follows from the independent sampling of Y_j and property (ii), and (11) follows by using the identity $\frac{1}{n}(1 - B_1)B_n = (1 - B_n)B_1$ and simplifying. The simplicity of (11) is a consequence of defining the shrinkage factor slightly different than Morris (1983a). Expression (11) can also be derived simply from regression formulae when covariates are not used by simplifying the expression $\text{Var}(Y_{n+1} | \bar{Y}_n) = \text{Var}(Y_{n+1}) - B_n^2 \text{Var}(Y_n) = \text{Var}(Y_{n+1}) \left(1 - B_n^2 \frac{\text{Var}(Y_n)}{\text{Var}(\bar{Y}_n)}\right)$.

Thus, the PEB rule gives Y_{n+1} a positive screen whenever

$$Y_{n+1} > \hat{\mu}_n + z_{1-f_0} \sqrt{1 - B_1B_n} \quad (12)$$

Note that when no history is available the PEB rule has $B_n = B_0 = 0$, and the PEB rule equals the ST rule. Subjects with long screening history have $B_\infty = 1$ and the PEB rule equals the limiting rule (7). Between these extremes the PEB rule provides a smooth generalization of these rules. A specific example of the PEB rule is given in Section 5.

The derivations above assume the population parameters are estimated with high precision. If the model parameters are estimated with substantial error then (11) must be adjusted to reflect this added uncertainty. Perhaps the simplest approach would be to add additional terms to (11) to inflate the variance until data becomes more abundant (for example see Section 5 of Morris, 1983b).

4. PERFORMANCE OF SCREENING RULES IN CANCER CASES

This section investigates the behavior of the PEB rule in diseased subjects by providing expressions for its sensitivity, or true positive rate (TPR). The exact TPR of any particular screening rule depends on how soon after cancer onset the screen takes place, and on aspects the tumors growth behavior which is not known. The TPR of the NS and PEB rules depend also on the abundance of screening history at the time of the screen. Because we wish to avoid stating a growth model we express the TPR of each rule when the marker has elevated a fixed but arbitrary amount $\delta > 0$; that is, the marker Y_{n+1} now has an expectation equal to $\mu + \delta$ instead of μ , but has the same variance.

With screening history summarized by \bar{y}_n and cancer onset occurring *after* the n^{th} the PEB screening rule with n historical screens is expressed as

$$t_n = P(Y_{n+1} > \hat{\mu} + z_{1-f_0} \sqrt{1 - B_1 B_n} \mid \bar{y}_n, E(Y_{n+1}) = \mu + \delta) \quad (13)$$

and from expression (11) this becomes

$$= P(Z > z_{1-f_0} - \delta / \sqrt{1 - B_1 B_n})$$

where Z represents a standard normal distribution, z_{1-f_0} its upper $1 - f_0$ quantile.

Moreover, by choosing $n = 0$ or $n = \infty$ expression (13) represents the TPR of the ST and limiting rules, respectively. Because $t_0 \leq t_n \leq t_\infty$ the PEB always out performs the ST rule and will eventually meet, but never exceed, the performance of the limiting rule. The PEB rule always out performs the NS rule too because $\sqrt{1 - B_1 B_n} \leq \sqrt{(1 - B_1)(1 + 1/n)}$.

The PEB true positive rate (13) is computed assuming that the cancer onset occurs following the n^{th} screen, which assures that screening history is uncontaminated with any false negative measurements. False negatives must be addressed because a marker that elevates but not enough to cause a positive screen will cause (12) to deviate from optimal on subsequent screens. It is not immediately apparent that the PEB rule should still be preferred to the ST rule in the presence of false negatives. Section 6 shows that individuals and the population should prefer the PEB rule to the ST rule even with the risk of a false negative. We first give an example of the PEB rule using the CA 125 ovarian cancer tumor marker.

5. EXAMPLE OF THE PEB RULE FOR OVARIAN CANCER

Most women diagnosed with ovarian cancer are diagnosed at a late stage, and nearly all die from their disease. However, most of those few diagnosed with early stage disease survive. This suggests that early detection of ovarian cancer may have a dramatic impact on mortality. Because the disease is relatively rare, it is difficult to obtain serial specimens collected during the suspected pre clinical period. However, many longitudinal cohort studies have very large stores of serum for asymptomatic women that can be used to estimate the PEB algorithm. For example, a ten year study of over 1,200 high risk (Karlan *et al.*, 1993) women was used by Crump *et al.* (2000) to determine the information sufficient to implement the PEB algorithm for five different ovarian cancer tumor markers. Here we present the algorithm for the CA 125 tumor marker.

5.1. Implementing the algorithm

Many studies report that $\log(\text{CA } 125)$ in healthy subjects has a normal distribution, with B_1 falling between 0.6 in a high-risk population (Crump *et al.*, 2000) and 0.9, in a general risk population (Pauler *et al.*, 2001). The population mean $\log(\text{CA } 125)$ level, $\bar{\mu}$ has been found to depend on the type of CA 125 assay used (Davelaar *et al.*, 1988), and the the menopausal status and race of the women (Crump *et al.*, 2000; Karlan *et al.*, 1993; Pauler *et al.*, 2001). However, the value of B_1 is unaffected by these covariates (e.g., pre and post menopausal women appear to have the same B_1 even though they have different $\bar{\mu}$ (Crump *et al.*, 2000)).

The PEB screening algorithm presented here uses a false positive rate $f_0 = 0.01$, and $B_1 = 0.6$, the lowest reported value B_1 for $\log(\text{CA } 125)$. Higher values of B_1 will imply better performance of the PEB rule compared to the ST rule. For convenience we will use the mean and marginal variance of $\log(\text{CA } 125)$ are selected as $\bar{\mu} = \log(10)$ and $\sqrt{\sigma^2 + \tau^2} = 0.5$, respectively. This mean is similar to that observed for a population of post-menopausal women (Crump *et al.*, 2000), and a different mean should be used if applied to a pre-menopausal population. Because serial correlation between subsequent Log (CA 125) values cannot be detected when observed more than one month apart (Crump *et al.*, 2000), we may safely ignore serial correlation in our model for any screening program with intervals exceeding that amount of time.

For a fixed amount of history, n , the PEB rule (12) linear in \bar{y}_n , with slope equaling B_n , and so it can be displayed graphically, as in Figure 1(a). The horizontal axis represents the mean of a woman's historical $\log(\text{CA } 125)$ levels, and each line in the figure is labeled by n , the number of screens used to compute the mean. The line with label $n = 0$ represents the ST ruled and has intercept $z_{1-f_0} = z_{0.99} = 2.33$, and slope $B_0 = 0$. Subsequent screens have slopes B_n that increase to $B_\infty = 1$, and an intercept that reduces to $2.33\sqrt{1-B_1}$. We can transform back to the raw CA 125 scale too, as in Figure 1(b), whose horizontal line gives a CA 125 threshold of $\exp(\bar{\mu} + z_{1-f_0} \times \sqrt{\sigma^2 + \tau^2}) = \exp(\log(10) + 2.33 \times 0.50) = 32$.

Figure 1(a) shows that the PEB rule diverges from the ST rule as history accumulates. The few women who have naturally high concentrations will have thresholds above the ST rule, but the majority will have lower thresholds, some dramatically so. Because an average healthy woman has $\bar{y}_n = 0$ (reduced model), we see that far more than half of all women will have a lower threshold if using the PEB rule. Indeed, average women have a ST rule threshold of 2.33, but becomes 80% of it (or 20% smaller) after only a single screen, $\sqrt{1-0.6^2} = 1.86$, and only 20% (or 80% smaller) of it after several screens, $\sqrt{1-0.6} = 1.47$.

Figure 1(a) shows all women having histories with \bar{y}_n less than approximately $\bar{y}_n = 1$ are given lower thresholds. The exact fraction with a lower threshold is given by $P(\bar{Y}_n < 1)$ which depends on the specific value of B_1 and n ; larger values of either increase the fraction with lowered thresholds. For our example 84%, 88%, 89%, and 95% of women are given lower thresholds for $n = 1, 2, 3, \infty$, respectively. Thus, when using

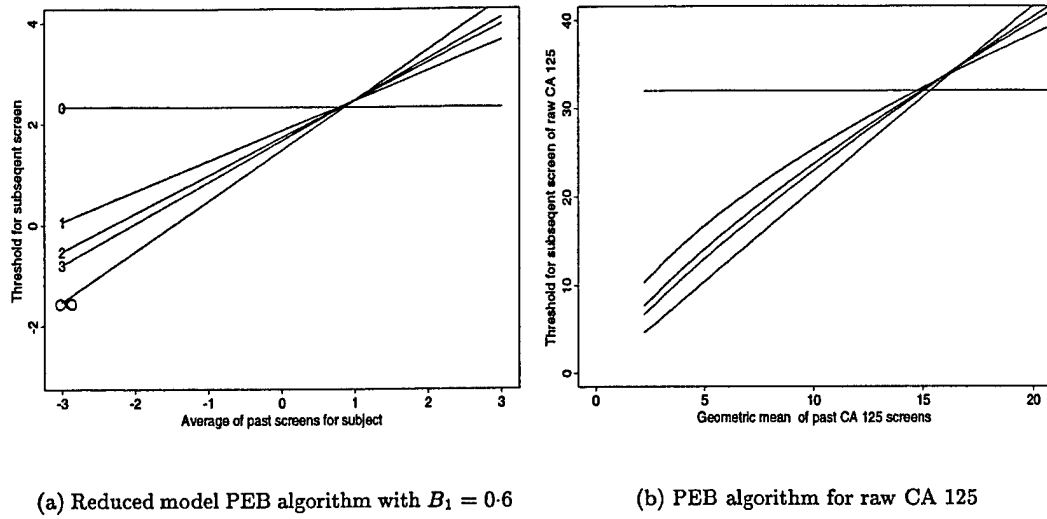


Fig. 1. Graphic representation of the PEB screening rules on the log scale (left) and raw scale (right) with false population and individual positive rates equal to $f_0 = 0.01$

the PEB rule the overwhelming majority of women will have a lower threshold than the ST rule gives them, and so their cancers can be detected earlier. A key point is that the PEB rule gives this lower threshold while maintaining the same population wide false positive rate.

Of course, lowering the threshold will increase an *individual's* false positive rate compared to the ST rule. In order to maintain the population-wide error rate the increase experienced by the majority must be offset by a concomitant decrease found by the minority. The fact that increasing the threshold in a small fraction of women permits drastic reductions in the thresholds of the majority is a consequence of the fact that when markers behave heterogeneously the ST thresholds are determined by a few extreme case. Indeed, in our example 5% of women experience over half the false positives when the ST rule is used. The PEB rule spreads the burden of false positive evenly throughout the population so that all women have the same burden, or better, can even permit each woman to choose her own FPR, depending on her own screening history. A consequence is a dramatic increase in the cancer detection ability, as we see next.

5.2. The behavior in cases

The TPR of the PEB rule to detect an elevation $\delta > 0$ is given by t_n , in (13). The practical significance of any elevation must be made on a case by case basis. For demonstration purposes only we choose $\delta = 1$

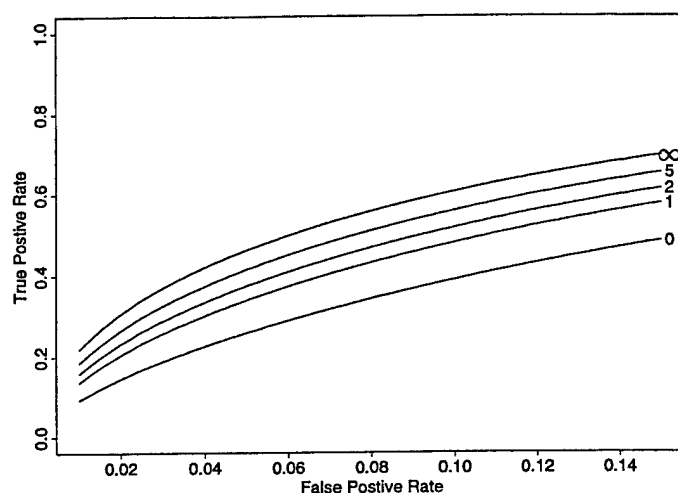


Fig. 2. Sequence of curves relating the false positive rate to the true positive rate of the PEB screening rule. Number next to the curve indicates the number of screens controlled for. Curve labelled 0 represents the ROC curve for the ST Rule.

as a convenient and arbitrary marker elevation of one marginal standard error. The solid curves in Figure 2 relates the TPR (vertical axis) versus FPR (horizontal axis) over a low range of FPR relevant for screening (i.e., low false positive rates). Such curves are often referred to as receiver-operating characteristic (ROC) curves and may be used to characterize the performance of diagnostic tests (Hanley and McNeil, 82). Each plot represents the ROC curve with different values of n .

When no history is available (Figure 2(a)) the PEB and ST rules detect 26% of the cases at $f_0 = 0.05$. When only a single historic screen is available nearly 34% of such cases can be detected by the PEB rule, and with one additional screen the PEB rule finds 38% of them. With substantial history ($n = \infty$) the PEB rule finds 47% of cases, which represents a relative 77% improvement over the ST rule. With $f_0 = 0.01$ these percentages are 10% 15% 17% and 23%, respectively. The gain in performance will be greater for larger values of B_1 , but for any fixed elevation, and any $B_1 > 0$ the PEB ROC curves will uniformly dominate the ST rules.

Note that for even this modest B_1 controlling for only a single historical screen ($n = 1$) achieves a 30% increase in TPR, which is nearly half the potentially achievable amount. This suggests that only a few recent screens need to be controlled for in order to obtain the majority of benefit of the PEB rule. Screens in the distant past will add only marginal benefit and may be ignored without substantial loss of performance.

5.3. The performance in a population

A comprehensive ovarian cancer microsimulation model has been used to evaluate the behavior of the PEB algorithm in a large cohort of women in the United States (McIntosh *et al.*, 2001; Urban *et al.*, 2001). The microsimulation model, explained in detail in by Urban *et al.* (1998) and Urban *et al.* (1997), is intended to examine the cost and health impact of ovarian cancer screening programs. These simulations are quite comprehensive. The life-histories (cancer incidence, cancer death, treatment, and and other cause death) for the US population are modeled to reflect US Census Bureau life tables and the age adjusted US cancer incidence.

Urban *et al.* (2001) used the microsimulation model and the PEB algorithm to estimate the performance of screening expected when using a screening protocol similar to that outlined by Skates and Singer (1991), who recommend two levels of positivity for a screen: positive screen for extreme elevations and early recall for modest ones. Note that for ovarian cancer a positive screen from a marker results in referral to ultrasound, not surgery. The PEB algorithm used a false positive rate of 2% to defined the extreme, and 15% for modest (i.e., 13% of healthy women were recalled early). With annual screening this configuration of the PEB algorithm detects 70% of cancers before their clinical diagnosis, whereas the ST rule finds only 46% of them. Importantly, the PEB rule finds over 50% when at early stage compared to mere 30% by the ST rule, and only 20% without screening. The PEB rule found an expected 31% drop in cause specific mortality compared to the 18% of the ST rule. This simulation study used $B_1 = 0.6$, which is lower than what is now reported in general risk populations with a modern assay (Skates and Pauler, 2001; Pauler *et al.*, 2001), and using higher values of B_1 would find even greater improvement.

The predictions of Urban *et al.* (2001) are very close to that predicted by the complex algorithm described by Skates and Pauler (2001): they predict 60% of cancers predicted at early stage. These simulations are not exactly comparable – Skates and Pauler (2001) assume a larger B_1 , and a longer duration of early stage disease – but their performance will agree even more closely if made more so because both of these differences will make the PEB algorithm perform even better than reported by Urban *et al.* (2001).

6. ROBUSTNESS OF THE PEB RULE TO FALSE NEGATIVES AND THE TUMOR GROWTH MODEL

Following a false negative the PEB threshold will be slightly too high because, although an elevation was not dramatic enough to cause a positive screen, on average false negative screens will be biased upward. Here we summarize formal, informal, and simulation evidence to argue that contamination is not a problem sufficient enough to prevent the strong preference of the PEB rule over the ST rule even for subjects who experience a false negative.

On a population level the PEB rule can be said to be superior to the ST rule if subjects can be expected, on average, to have a better outcome under the PEB rule than the ST rule. This is the criteria used by Skates and Pauler (2001) to determine the superiority of their algorithms. The simulation studies of Urban *et al.* (2001), summarized in Section 5.3, showed that the PEB rule too dominates the ST rule in all outcome categories when using CA 125 to screen ovarian cancer. McIntosh *et al.* (2001) used the same model to show that the dominance holds for a wide variety of other markers by varying the range of B_1 between 0.1 and 0.9.

Importantly, of the over 15,000 cancer cases simulated by McIntosh *et al.* (2001) not a single subject was diagnosed later using the PEB rule than using the ST rule, even though false negative screens occurred frequently. This suggests that the PEB rule dominance may extend beyond the population to the individual level. We may use to simple framework of the PEB rule to investigate this more formally. Consider a subject whose initial screen is falsely negative, and is elevated by δ_1 , second screen by δ_2 , etc.. The $n + 1^{th}$ screen is elevated by δ_{n+1} and the PEB threshold is biased upward by $\bar{\delta}_n B_n$ where $\bar{\delta}_n = \sum_{j=1}^n \delta_j / n$. Any marker growth model where the difference $\delta_{n+1} - \bar{\delta}_n B_n$ increases with time assure that the detection rate of the PEB algorithm will not degrade following a false negative because then increase in the marker growth outruns the bias in the PEB estimate. A convex growth model (e.g., linear growth) on the transformed scale is sufficient for this, but not necessary. If not all the historical screens are false negative then the effect of a false negative is dampened, with the upward bias equaling $\frac{c}{n} \bar{\delta}_c B_n$, where $c \leq n$ represents the number of n screens that are false negatives.

We can make even more precise statements if we assume a linear growth model with rate β per screening interval. The upward bias in the threshold is then less than $\frac{c}{n} \frac{\beta c}{2} B_n$, and so a true elevation of $\beta(c + 1)$ appears to the PEB rule as being too small due to contamination by a factor $1 - \frac{c}{c+1} \frac{c}{2n} B_n$. However, recall that the PEB rule can detect elevations that are $\sqrt{1 - B_1 B_n}$ smaller than the ST rule. Thus, whenever $1 - \frac{c}{c+1} \frac{c}{2n} B_n > \sqrt{1 - B_1 B_n}$ then the ROC curve for the PEB rule dominates that of the ST rule. The inequality simplifies to

$$B_1 > \left(\frac{c}{c+1} \right) \left(\frac{c}{n} \right) \left(1 - \left(\frac{c}{c+1} \right) \left(\frac{c}{n} \right) \frac{B_n}{4} \right)$$

and when it holds a very strong statement can be made: the PEB rule has higher detection than the ST rule following c false negatives. For example, subjects whose only screen were false negative ($n = 1, c = 1$) then any $B_1 > 17/32$ satisfies the inequality. Far lower B_1 are sufficient if the history contained only a single true negative result. Note that the expression above is independent of the growth parameter β , and so a linear or greater growth model, on the transformed scale, whatever the parameter or growth shape, is sufficient for the PEB rule to improve over the ST rule. This is in contrast to the complex change point models which

rely on a specific growth curve and parameters for them to claim optimal performance. It is unclear how those algorithms will perform with models or parameters different than assumed.

These conclusion are stronger than the population based ones demonstrated by the microsimulation models: even if the inequality does not hold the population can still find greater performance because the degradation following a false negative can be offset by the lower probability of having one in the first place. However, the inequality above does show that the only subjects at risk of later detection with a sequential rule are those who experience a false negative at the start of the screening program. If B_1 is very small and early false negatives are thought to be common, then diligence may be used early on, but after only a few screens, the PEB rule dominates the ST rule for even that critical sub-population.

7. SUMMARY AND CONCLUSIONS

Our approach uses Parametric Empirical Bayes methods to control for screening history and give each subject a pre specified false positive rate at every screen. Although the PEB rule can maintain the same population wide false positive rate as the ST rule, the great majority of subjects will have lower thresholds, and so their cancers can be detected earlier. The PEB algorithm works best, compared to the ST rule, with markers that have high population heterogeneity. The PEB rule has advantage over other complex methods because by comparison it has modest data needs that are modest, and does not depend on a specific growth model.

The PEB algorithm may be particularly useful when natural history is not well known, but it is important to consider what may be lost by using it instead of algorithms like Skates and Pauler (2001) when the pre clinical model is well known. Complex algorithms may have particular advantage when differential screening intervals are used. This is due to their explicit representation of time in their approach. For example, Skates and Pauler treat a marker elevation as more significant if it occurs over a shorter time interval, whereas the PEB method considers only the magnitude of the elevation. Although this may not be important in a screening program with a fixed screening interval, it may be with screening protocols similar to that recommended by Skates and Pauler (2001), who schedule the screening interval based on observed marker changes.

As appealing as accommodating time is, doing so apparently requires stating a complex change point model which, as we have argued, is difficult to estimate. When their model is correct, theory suggests that the Skates and Pauler approach will perform better than the PEB method, but the simulation results summarized in Section 5.3 (which use differential recall rates comparable to Skates and Pauler) suggest the gain may be only small, at least for ovarian cancer and CA 125. It is also unclear how much their algorithm will degrade with model misspecification. At this time no general statement can be made about the relative

performance of the PEB method and the more complex methods.

Several modifications to the PEB rule may be considered to make them even more robust and widely applicable. For example, the PEB rules above use the complete screening history. However, screening takes place over several years, or decades, and subjects will accumulate an extraordinary screening history in their lifetimes. Because most of the gain from longitudinal screening can come with only a few historical screens, depending on the value of B_1 (see Section 5), a practical adaptation of the PEB method is to omit the distant past. This will result in little degradation of screening performance but will decrease dependence on the model for healthy subject's trajectories.

The PEB method assumes that the markers in the healthy subjects can, with some transformation, be represented in a hierarchical normal model. The existence of a transformation to marginal normality is automatic for any marker measured on a continuous scale. Only the addition of the within subject normality, with equal variance, must be accommodated within the parametric structure. If that assumption appears to fail, the PEB methods presented here can be extended, almost without modification, to a wider class of continuous and discrete distributions, including the Gamma, Poisson, and binomial families (Morris, 1982, 1983a). The assumptions of the general PEB approach are likely appropriate for a large number of tumor markers, and assist timely translation of tumor marker discoveries.

Acknowledgements

Supported by National Cancer Institute grants NCI 1 P50 CA83636 and NCI 5 R01 CA75494. We would like to acknowledge many valuable discussions of Carl Morris, Beth Karlan, Steve Skates and Donna Pauler, but do not wish to imply their endorsement of the manuscript's content.

REFERENCES

- Baron, A., Lafky, J., Boardman, C., Balasubramaniam, S., Suman, V., Podratz, K., and Maihle, N. (1999). Serum serbb1 and epidermal growth factor levels as tumor biomarkers in women with stage iii or iv epithelial ovarian cancer. *Cancer Epidemiology, Biomarkers, and Prevention* 8, 129–137.
- Bast, R. C., Siegal, F. P., Runowicz, C., and Klug, T. L. -. (1985). Elevation of serum ca 125 prior to diagnosis of an epithelial ovarian carcinoma. *Gynecologic Oncology* 22, 115–120.
- Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician* 39, 83–87.
- Crump, K. C., McIntosh, M. W., Urban, N., Anderson, G., and Karlan, B. Y. (2000). Ovarian cancer tumor marker behavior in asymptomatic healthy women: Implications for screening. *Cancer Epidemiology, Biomarkers and Prevention* 9, 1107.
- Davelaar, E., van Kamp, G., and Verstraeten, R.A. and Kenemans, P. (1988). Comparison of seven immunoassays for the quantification of ca125 antigen in serum. *Clinical Chemistry* 44, 1417–1422.
- Efron, B. and Morris, C. N. (1997). Stein's paradox in statistics. *Scientific American* 236, 5, 119–127.
- Hanley, J. A. and McNeil, B. J. (82). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143, 29–36.
- Jacobs, I., Skates, S., MacDonald, N., Menon, U., Rosenthal, A., Davies, A., Woolas, R., Jeyarajah, A., Sibley, K., Lowe, D., and DH, O. (1999). Screening for ovarian cancer: a pilot randomised controlled trial. *Lancet* 353, 9177, 1207–1210.
- Karlan, B. Y., Raffel, L. J., Crvenkovic, G., Smrt, C., Chen, M. D., Lopez, E., Walla, C. A., Garber, C., Cane, P., and Sarti, D. A. (1993). A multidisciplinary approach to the early detection of ovarian carcinoma: rationale, protocol design, and early results. *American Journal of Obstetrics & Gynecology* 169, 3, 494–501.
- McIntosh, M., Urban, N., and Karlan, B. (2001). Screening algorithms for novel tumor markers. *Accepted to: Cancer Biomarkers, Epidemiology, and Prevention* .
- McIntosh, M. W. and Pepe, M. (2001). Optimal rules for combining tumor markers for cancer screening. *Accepted to Biometrics* .
- Morrell, C. H., Pearson, J. D., Carter, H. B., and Brant, L. J. (1995). Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer. *Journal of the American Statistical Association* 90, 45–53.

- Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics* **10**, 65–80.
- Morris, C. N. (1983a). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics* **11**, 515–529.
- Morris, C. N. (1983b). Parametric empirical Bayes inference: Theory and applications (c/r: P55–65). *Journal of the American Statistical Association* **78**, 47–55.
- Pauler, D. K., Menon, U., McIntosh, M. W., Symecko, H. L., and Skates, S. J. (2001). Factors influencing serum ca 125ii levels in healthy postmenopausal women. *Cancer Epidemiology, Biomarkers and Prevention* **10**, 489–493.
- Skates, C. J. and Pauler, D. K. (2001). Screening based on the risk of cancer calculation from bayesian hierarchical change-point models of longitudinal markers. *Journal of the American Statistical Association* **96**, 429–439.
- Skates, S. and Singer, D. (1991). Quantifying the potential benefit of ca 125 screening for ovarian cancer. *Journal of Clinical Epidemiology* **44**, 365–380.
- Slate, E. and Clark, L. (1999). *Case Studies in Bayesian Statistics IV*, vol. IV, chap. An application of Bayesian retrospective and prospective changepoint identification, 511–534. New York: Springer Verlag.
- Slate, E. and Cronin, K. A. (1997). *Case Studies in Bayesian Statistics IV*, vol. III, chap. Changepoint Modeling of Longitudinal PSA as a Biomarker for Prostate Cancer, 435–456. New York: Springer Verlag.
- Urban, N., Drescher, C., Clarke, L., and Kiviat, N. (1998). *Socioeconomics of Ovarian Cancer Screening*, chap. Ovarian Cancer. Oxford Isis Medical Media.
- Urban, N., Drescher, C., Etzioni, R., and Colby, C. (1997). Use of a stochastic simulation model to identify an efficient protocol for ovarian cancer screening. *Controlled Clinical Trials* **18**, 251–270.
- Urban, N., McIntosh, M., Clarke, L., Jacobs, I., Karlan, B., Anderson, G., and Drescher, C. (2001). *Socioeconomics of Ovarian Cancer Screening*, chap. Ovarian Cancer. Oxford University Press.
- Zurawski, V. R., Orjaster, H., Andersen, A., and Jellum, E. (1988). Elevated serum ca 125 levels prior to diagnosis of ovarian neoplasia: Relevance for early detection of ovarian cancer. *International. Journal of Cancer* **42**, 677–680.

[Received]

Title: Generating Longitudinal Cancer Screening Algorithms for Novel Tumor Markers

Running Title: Screening algorithms for novel markers

Martin W. McIntosh, Ph.D, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue MP-900, Seattle Washington, 98109-1024, email mmcintos@fhcrc.org, phone: (206)667-4612, Fax: (206)667-7850.

Nicole Urban, Sc. D, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue MP-900, Seattle Washington, 98109-1024, email nurban@fhcrc.org, phone: (206)667-4677, Fax: (206)667-7850.

Beth Karlan, M.D., Director, Division of Gynecologic Oncology, Cedars-Sinai Medical Center and Professor of Obstetrics and Gynecology, UCLA School of Medicine, 8700 Beverly Boulevard, #160W, Los Angeles, California 90048, email: karlanb@csbs.org, phone: (310) 423-3302; Fax: (310) 423-0140

Contact author: Martin McIntosh.

Number of: Text pages 16, Tables 2, Figures 1

Abstract

Objective: Recent advances in molecular technology are leading to the discovery of new tumors markers which may be useful for cancer screening and early diagnosis. Translating a potential tumor marker from the laboratory to its use in patient care may require an algorithm, or screening rule, for its application. An algorithm that can detect the smallest deviation from a defined norm is likely to achieve the highest sensitivity, but any practical screening algorithm must do so with strict controls on test specificity to avoid false positive results and unnecessary patient alarm and risk. Longitudinal algorithms, that make use of prior tumor marker values and trends, are likely to obtain improvements over single threshold rules. Thus far, a few longitudinal screening algorithms have been proposed (for example, using serial PSA values for the detection of prostate cancer, and serial CA125 values for the detection of ovarian cancer), but these algorithms are not appropriate for novel tumor marker discoveries because they rely on assumptions that may not translate to the new marker's behavior. The algorithm presented here is motivated by 1) the need to develop an algorithm for early detection using novel markers, 2) the practical demands on data and specimen availability, and 3) the need to be robust enough to accommodate a wide range of tumor growth behavior. **Methods:** We use Parametric Empirical Bayes (PEB) statistical theory to model the trajectory of markers over time in a cohort of asymptomatic healthy subjects, and use the estimated trajectory to produce person specific thresholds that depend on each person's screening history. The thresholds are chosen to give the person (or population) a specified false positive rate. **Results:** The resulting algorithm is simple, and can be represented in a simple graph, or a chart. The statistical analysis needed to generate the algorithm can be found in nearly every basic statistical package. The algorithm is highly robust, and can detect a wide range of tumor behaviors **Conclusions:** The PEB screening algorithm should take a central role when evaluating marker discoveries for use in screening. The algorithm is particularly useful when screening with a new marker whose behavior in the pre clinical period is not well known.

Introduction

The most common tumor marker screening algorithm is the single threshold (ST) rule which gives a positive (likely to have disease) outcome for any subject whose marker concentration deviates from a common population-wide threshold. When tumor marker behavior is heterogeneous across individuals however, a better approach is to use a longitudinal algorithm to tailor the threshold to the individual's own screening history. Longitudinal algorithms, such as that using PSA values to detect prostate cancer¹⁻³ or CA125 values to diagnose ovarian cancer,⁴ have been proposed, but those algorithms are not appropriate for general use, for new tumor marker discoveries, or for other diseases. These algorithms are computationally intensive, and require substantial modeling of the tumor marker growth during the pre-clinical period. When modeled adequately, these algorithms yield sensitivity, but the specimens needed to observe and define the pre-clinical trajectory are so rare that these approaches are impractical for general use. The marker trajectory can be observed only by having access to several serially collected specimens taken from several cases during their pre-clinical period. Such observations are extremely rare and for many diseases and not even stored blood specimens collected as part of large longitudinal studies may be sufficient. For example, the ovarian cancer screening study by Jacobs⁵ randomized half of nearly 22,000 postmenopausal women to three annual primary screens with CA125 and secondary screen with ultrasound. Only six ovarian cancer cases were found during screening, with another ten reported within a seven-year follow-up. However most of cases (the ten found during followup) will not have had serial serum samples taken during their early pre clinical period. Specimens from the cases found during screening were collected at annual intervals, not frequent enough to observe the early pre clinical behavior of a disease such as ovarian cancer where the median pre-clinical duration is thought to be under 18-months. Moreover, in diseases where stored specimens are sufficient, the storage conditions cannot anticipate the processing requirements of all future tumor markers.

An alternative to those data- and computation- intensive algorithms is presented here. This algorithm is particularly appropriate for screening with new markers or in new populations. Our approach generates

screening rules by focusing its attention on the trajectory of the markers in healthy women, for whom data are plentiful. The algorithm then determines if a general, rather than a specific (i.e., exponential or linear) change in marker behavior has occurred. The algorithm may be calibrated to work using any specificity, chosen either for the population or for each individual. The required specificity, along with a subject's screening history, are used together to compute a person specific threshold that is used to assess the next screening event. In this way the marker's behavior is characterized and used to reflect the disease process or tumor growth and progression. Computing the threshold is very simple, and involves computing the sample average of subjects screening history. Then, the screening threshold can be read off of a graph or looked up in a simple table. Our approach is based on Parametric Empirical Bayes (PEB) statistical methods, and we refer to our method simply as the PEB screening algorithm or simply the PEB rule. A full technical description of the algorithm derivation and theoretical properties is given by McIntosh & Urban ⁶, and is outlined below enough for the reader to understand its properties and implement it.

The PEB screening rule uses PEB statistical theory to infer each subject's individual normal marker levels accounting for the population heterogeneity. The individuals' estimated normal levels are then used to produce a threshold tailored to his/her normal behavior. This normal behavior is revealed over time by event-free screening. The PEB screening rule has the surprising property of giving the vast majority of all women a threshold far lower than the single threshold rule but maintains the same population wide specificity. Thus, in theory, cancers can be detected at an earlier stage, without increasing the burden on the healthy population. In can be shown technically that the PEB rule uniformly outperforms the single threshold rule in that for *any* level of specificity the PEB rule achieves a higher sensitivity ⁶.

The following two sections describe the behavior the PEB algorithm requires of a tumor marker and the data needed to implement the algorithm. We also contrast the PEB requirements with the existing longitudinal algorithms for CA125 and PSA mentioned above. We do this to point out that that the PEB assumptions are

typically less restrictive, but also to point out that those other approaches may be considered when their more stringent requirements are met. The third section below describes how to estimate the parameters needed for the PEB algorithm, and presents the algebraic expression to determine the threshold. An informal discussion of the technical aspects of the PEB algorithm follows. Finally, we discuss aspects of implementing the PEB rule that should be considered before applying the model to an actual population.

Methods

Marker behavior required of the PEB screening algorithm

A PEB screening algorithm can be generated for any marker meeting a modest set of requirements. Specifically, 1) marker concentrations in healthy subjects are continuous, 2) marker concentrations tend to either remain unchanged or change in a known direction (elevate or decrease) with the onset of cancer, or other pathologic/disorder, and 3) covariates that describe large differences in marker behavior in healthy subjects are available, and 4) the serial variation of the marker is not extreme over time in healthy subjects. These assumptions appear widely applicable to many tumor markers.^{7, 8} Importantly, note that no specific assumptions are made about the growth of a marker at cancer onset other than it elevates (or declines) with cellular transformation.

The first assumption suggests that the marker concentrations will have a normal distribution in healthy subjects. The second assumption allows us to look for deviation, in the marker concentration in a known direction, when disease develops. The final two assumptions allow us to model the trajectory over time for healthy women and without substantial mathematical modeling. Each of these assumptions can be relaxed for the general PEB rule, but not for the simpler form presented here for potential clinical application to early cancer detection. (see McIntosh & Urban⁶).

The previous longitudinal algorithms also require these assumptions but in those applications they are even more specific and restrictive. Those algorithms require a parametric (i.e., log) transformation to normality (restrictive form of 1), a constant linear marker growth after cancer onset (restrictive form of 2), and have the same restrictions with serial correlation. Moreover, the transformation to normality must be on the same scale that gives the marker a linear growth. These restrictions are violated, for example, with any marker that is normal on the raw scale, with an additive exponential growth model. It is also violated if the tumor growth model meets all assumptions but the growth is not constant over the entire pre-clinical period. There is no evidence that the same uniform growth model should hold for every tumor type⁹ much less for every potential marker, whose growth may not follow the tumor volume exactly. However, if those requirements are met, and the data are available to estimate it, those complex algorithms may find higher sensitivity than the PEB algorithm presented here.

Specimen requirements to implement the PEB algorithm

Novel tumor markers are typically evaluated by measuring their concentrations in stored specimens. Most commonly, the stored specimens are of two types; specimens from healthy control subjects and specimens from clinically presenting cases. A second easily obtained category of specimens are serially collected samples, measured in the same control individual at two time points, typically separated by several months. These serially measured samples from control subjects are necessary to implement the PEB algorithm. Sequential measurements on healthy controls, while not as abundant as cross-sectional specimens from healthy controls, can be obtained in substantial abundance over a short period of time. For technical and practical reasons, the duration between specimen collections should not be too frequent so that the serial correlation assumption above is violated, and not more infrequent than the proposed screening interval. The duration between intervals is likely to be both marker and disease dependent, but a large set of markers for ovarian cancer have found that specimens taken at one month intervals were sufficiently spaced to satisfy these requirements.

The PEB algorithm can be further improved if individual characteristics that can predict the marker behavior in healthy subjects are known. For example, the ovarian cancer tumor marker CA125 is known to have substantial differences based on a woman's menopausal status or the and presence of benign gynecologic conditions. These conditions may not contribute to the ovarian cancer risk substantially, but they may cause statistically significant changes in the CA125 values.¹⁰ Characteristics believed to influence the behavior of the markers in healthy subjects should be recorded, and evaluated before implementing the PEB algorithm. The algorithm presented below will assume it is estimated separately for women having similar values of such covariates, but extensions can be made to include the covariates in the PEB rule formally (see McIntosh & Urban ⁶). Omitting such variables when estimating the PEB rule will not invalidate the algorithm, but it will be less efficient. This point is true for all screening algorithms, not only the PEB approach. The precision of the PEB algorithm will depend on the number of serial controls available, but only a modest number are needed to implement the algorithm

The specimen needs of the PEB algorithm are far less demanding than those required by the more complex longitudinal screening approaches. Those algorithms have great demands on the rarest type of specimen: serial observations in cancer cases obtained at high frequency during the pre-clinical period. Such specimens are extremely difficult to come by, and typically suggest mean that implementing the algorithms for new markers or diseases would rely on untested assumptions. Thus, to use those methods for all but highly prevalent and slow growing diseases requires, like the early work by Skates ^{10, 11}, a substantial amount of time to accumulate the information in order to fit and implement the algorithm.⁴

How to estimate the parameters and to implement the PEB algorithm

Presently, we assume that the marker concentration has been measured serially on a large number, N , of healthy control subjects. For clarity exposition we assume (a) the control subjects have contributed exactly two marker observations, (b) all subjects are relatively homogenous with regard to covariates than predict the normal marker levels (see section above), and (c) the markers are measured on a scale giving them a normal distribution in

controls. The first and second assumptions are only made to reduce notation burden, and in practice are not needed (see Pauler et al ¹² for details about how to measure these parameters in a more general setting and McIntosh and Urban ⁶ on how to extend those to include covariates more formally). The final assumption (c) is necessary for our method to function properly, but it does not limit the generality of our method for any marker having continuous measurements because then the existence of a transformation to normality is assured. Transformation can be done by some mathematical transformation such as log, or by simply translating the quantiles of the empirical marker distribution to that of the standard normal distribution.

The PEB algorithm requires that we examine the marker concentration mean, denoted $\bar{\mu}$, and decompose its variance into its total, within, and between components, denoted by V, σ^2, τ^2 respectively. Note that by definition $V = \sigma^2 + \tau^2$, and so in actuality we need to estimate only three parameters. If subject i donates two sequential marker concentrations we denote the first and second level with Y_{i1} and Y_{i2} , respectively. Our parameters are defined formally as follows: $\bar{\mu} = \text{mean}[Y_{ij}]$, the mean of all marker measurements, $V = \text{Var}[Y_{ij}]$, the variance of the markers in the population, $\sigma^2 = \frac{1}{2} \text{Var}[Y_{i2} - Y_{i1}]$ the marker variance *within* a subject, and $\tau^2 = V - \sigma^2$, the between subject *heterogeneity* of the markers. Decomposing the marker variance into its within- and between-components allows us to characterize the *relative* subject heterogeneity in the population by the quantity $B_1 = \frac{\tau^2}{V}$, which is often called the intra-class correlation (the reason for the subscript will become apparent below). Describing the most efficient manner to estimate these quantities is not in the scope of this article, but the expressions above do suggest simple ways to do so (see Pauler et al ¹² for more efficient methods that can accommodate different numbers of observations per subject).

The PEB rule uses these quantities to determine the threshold for a subject from his/her screening history.

Suppose a subject with n historical screens with an average concentration denoted by \bar{y}_n is about to undergo

her next screen, and we wish the screen to operate at specificity α . The threshold for the next screen is given by the PEB algorithm to be equal to:

$$\text{Threshold} = \mu + (\bar{y} - \mu)B_n + z_\alpha \sqrt{1 - B_1 B_n} \sqrt{V}$$

where

$$B_n = \frac{\tau^2}{\sigma^2/n + \tau^2}$$

and z_α is the α quantile of a standard normal distribution (i.e., $z_\alpha = 1.96$ when $\alpha = 0.975$). Note that the expression above only the mean \bar{y} and *shrinkage factor* B_n need to be recomputed from time to time, as all other quantities remain the same throughout the screening study. The PEB rule can also be represented in a graph or in a table, thus requiring that we only compute an average to determine the threshold. We demonstrate this below.

Results

Example: using CA 125 to screen a population of high-risk women

Here we demonstrate the PEB rule when using CA 125 to screen a population of women for ovarian cancer, and implement the algorithm with a specificity of 98%. Several publications have found that $\log(\text{CA 125})$, to the base e, has a highly normally distributed distribution^{4, 8, 10-12}, and so here we generate the PEB rule using $\log(\text{CA 125})$. The behavior of $\log(\text{CA 125})$ may depend heavily on the population and the type of assay used¹³. Here we use the population investigated by Crump et al.⁸, which is a population of women at high risk for ovarian cancer, and screened as part of an ongoing screening program at the Gilda Radner Ovarian Cancer Detection Program at Cedars-Sinai Medical Center in Los Angeles, California¹⁴. The following behavior of $\log(\text{CA 125})$ has been estimated in that population: $\mu = 2.27$, $V = 0.30$, $z_{0.98} = 2.05$, $\tau^2 = 0.21$, $\sigma^2 = 0.09$, and

$$B = \frac{0.21}{0.30} = 0.68.$$

Figure 1 represents the PEB screening rule graphically, and Table 1 shows the PEB rule in tabular form. For convenience, either can be used in place of the formal PEB expression above. Figure 1(a) represents the screening rule using $\log(\text{CA } 125)$ and Figure 1(b) transforms the PEB rule back to the raw scale. Note that the PEB rule is linear only on the transformed scale. We note that when summarizing the screening history for use on the raw scale (Figure 1(b)) the *geometric average* of the historic CA 125 concentrations are used, and not their numerical average. The numeric averages will tend to over estimate the geometric mean.

Table 1 here

The lines in Figure 1 show the serum concentration thresholds leading to a positive screen for a woman based on her screening history. The horizontal axes in Figure 1(a) represents the numeric average of her historic $\log(\text{CA } 125)$, and each line in the figure is used to determine the threshold based on this average and n , the number of historic screens used to compute the average. To determine the threshold for a woman undergoing screening, a clinician will

1. Compute the average of her previous $\log(\text{CA } 125)$ values.
2. Locate the line in graph, or the table column, corresponding to the number of historical screens used in step 1.
3. If using the table, simply read the threshold. For the graph, find the point on that line directly above the value computed in step 1, and then read directly across to the vertical axis to find the threshold.

The graph shown in Figure 1(b) works similarly, except that the horizontal axis represents the geometric average CA125 levels and the vertical axis gives the threshold on the raw scale. When on the raw scale, it is most convenient to have the PEB screening rule represented as in Table 1. Table 1 shows the threshold for a woman

based on her geometric average CA125 levels. The following example demonstrates how the PEB rule works on a hypothetical woman's screening history.

Figure 1 here

Consider a woman whose first three years experience with screening found CA 125 concentrations equal to 8, 16, and 12, in her first, second and third screen respectively. The natural log of these concentrations are 2.07, 2.77, and 2.48 respectively. At the initial screen, when no screening history is available, the PEB rule is mathematically equivalent to the single threshold rule, and so a threshold of $\text{Log}(30 \text{ U/ml})=3.40$ is used. The horizontal lines in Figure 1 represent this rule. The first screen result of 2.07, or 8 on the raw scale, is used to determine the threshold to evaluate her second screen. The second column of 2 of Table 1 shows that the second screen should use a threshold of 21.34 on the raw scale. Note that this threshold is far lower than the single threshold rule and initial screen. The third screen uses the average of the first two screens $2.42=(2.07+2.77)/2$, or their geometric average $\text{Exp}(2.42)=11.2$ to compute the next threshold. Table 1 shows the third screens should use a threshold of 24.65 U/ml. At the fourth screen, where the average history equals $2.44=(2.07+2.77+2.48)/3$ and with geometric average $11.5=\text{Exp}(2.44)$, Table 1 shows that a woman with a threshold above 25.87 should get a positive screen.

Estimating the performance of the PEB algorithm in a population

A comprehensive ovarian cancer microsimulation model has been used to evaluate the behavior of the PEB algorithm in a large cohort of women in the United States¹⁵. The microsimulation model ages 1 million women from the age of 50 until their death (including US cancer incidence, cancer death, treatment effects, and other causes death), then predicts their cancer outcomes after imposing a screening algorithm on the life histories. The microsimulation model, explained in detail in by Urban et al.¹⁶, and described briefly at the end of this section.

Urban et al.¹⁵ used the microsimulation model to predict the behavior of the PEB algorithm when used in a manner similar to that recommended by¹¹. They recommend two levels of positivity for a screen: positive screen for extreme elevations and early recall for modest ones. The PEB algorithm used here chose a positive rate of 2% to define the extreme, and 15% for modest (i.e., 13% of healthy women were recalled early). With annual screening this configuration of the PEB algorithm detects 70% of cancers before their clinical diagnosis, whereas the ST rule finds only 46% of them. Importantly, the PEB rule finds over 50% when at early stage compared to mere 30% by the ST rule, and only 20% without screening. The PEB rule found an expected 31% drop in cause specific mortality compared to the 18% of the ST rule. This simulation study used $B_1 = 0.6$, which is lower than what is now reported in general risk populations with modern assays¹². Using greater value would improve performance.

The predictions of Urban et al. are very close to that predicted by Skates and Pauler⁴, who predict 60% of cancers predicted at early stage. These simulations are not exactly comparable, and they will likely come to closer agreement if made more: Skates and Pauler assume a larger B_1 , a longer duration of early stage disease, and recall 25% of women early. Other aspects of their simulation are not described with enough detail to make the comparison exact. Nonetheless, the simplicity of the PEB rule, at least for CA 125, appears to be comparable to those more complex approaches that require far greater amounts of information, and computation.

The simulations summarized above are specific for ovarian cancer and CA 125. Here we use a comprehensive ovarian cancer microsimulation model to explore the performance of the PEB rule for hypothetical novel ovarian tumor markers having B_1 between 0 and 0.9. This is somewhat representative of the range found for a host of ovarian cancer markers⁸.

Table 2 summarizes the simulations when using a marker behaving like CA 125 in all respects but with widely different B_1 . Note that when $B_1 = 0$ the PEB rule equals the ST rule. The simulation ages 1 million women from age 50 to death with cancer incidence matching SEER. A total of 15,660 of the 1 million women are found with cancer before their natural death (nearly 1 in 64), and 77% of them die from their disease. The rows of Table 2 summarize their expected experience if they had instead been screened between ages 50 and 80 using the PEB algorithm. Note that the PEB algorithm used here is operating at a specificity of 99% and does not perform early recall. The results of Table 2 should not be compared to those by Skates and Pauler, because those perform early recall and operate with specificity as low as 75%.

Row 1 measures the fraction of all cases who would have been detected earlier, row 2 measures the average number of years earlier all cases would have been detected, row 3 represents the fraction of cases who would have been detected at an earlier stage, row 4 represents the absolute reduction (from 77%) of cancer cases whose cause of death would be other than ovarian cancer, and row 5 approximates the mean years extended for all cases. We emphasize that the performance in Table 2 averages the experience of all cases, not just those who were screen detected.

Table 2 here

The performance of screening improves as B_1 increases, and indeed shows that the PEB rule still achieves better performance than the ST rule even when B_1 is small, when the possibility of contamination is high. The ST rule detects 46% of the cases early, but the PEB rule can detect up to 50% of cases when B_1 is high (a relative 8% improvement). This improvement may seem trivial but cases detected by the ST rule may find detection occurring earlier with the PEB rule. The PEB rule detects cases on average up to 1 ½ months earlier than the ST

rule, (0.93 years versus 0.80 years), which for many women is large enough to achieve a stage shift. Indeed, only 35% have a stage shift with the ST rule, but up to 42% find one with the PEB rule (a 20% increase). Because mortality predictions depend primarily on a stage shift, this leads to a substantial improvement in cancer mortality; from 19% up to 26% (a 40% increase). We may also draw the conclusion that the PEB algorithm is more cost effective, because the algorithm achieves its performance without increasing the number of screening events.

Summary of the Ovarian Cancer Microsimulation model

For ovarian cancer and CA 125, screening is done in a multi-modal manner where CA 125 is the first stage of a screening procedure: a positive CA 125 screen is used to refer women to further diagnostic work up using Transvaginal Sonography (TVS). All women are assigned age at clinical diagnosis with ovarian cancer, stage of the disease at clinical diagnosis, and age at death due to ovarian cancer to match United States Cancer Incidence (SEER cancer incidence). We also assign each subject an age of death due to other causes to match the United States Census Bureau life tables. The microsimulation model then infers a pre clinical duration backward in time from the age of clinical diagnosis to its cancer onset. The pre clinical duration is assigned a value that depends on the woman's age and the cancer's stage at the time of detection. The cancer marker is then assigned growth behavior over the pre clinical period in a manner similar Skates and Pauler⁴ and Pauler et al¹²; $\log(\text{CA } 125)$ has linear growth with random (person specific) slope. The cancer is given four sequential stages over the pre clinical period representing four progressive SEER cancer stages, and an average of one fourth of their time is spent in each stage. Thus, the observable quantities for women when not screened match the United States SEER registry and US Census tables, and their unobserved quantities match currently understood behavior of the tumor biology.

An annual screening regimen is imposed on these women's life histories. We calibrate the PEB algorithm to operate at $\text{FPR}=0.01$, and the TVS component of the screen is assumed 100% sensitive for all tumors older than

6 months, and 80% sensitive for younger tumors^{18, 19}. If detected earlier by screening, the simulation model uses the age and stage at the screen detection to re-establish the woman's cause specific life expectancy. The primary benefit screening comes when a stage shift is achieved. Women diagnosed at an earlier stage are given extended cause specific life expectancy, and women without a stage shift are not given any benefit from screening. If the new ovarian cancer age of death extends past her assigned other cause of death then she does not die of the disease, and cancer mortality is reduced. Thus women can be unaffected by screening (if not detected or detected still at advanced stage), have life extended but still die of the disease (if new ovarian cancer death does not extend past other cause of death), or have life extended and die of another cause. Only in this later case has ovarian cancer specific mortality been reduced. The model compares the life histories under screening and without screening, and compares the rates of early detection, stage shifts, lead-time, and mortality effects.

Discussion

The usual method for screening with a tumor marker, the single threshold rule, gives all women the same population-wide threshold for positivity. This threshold is determined by finding a population wide reference range, or quantile, and defining a positive screen when the marker exceeds that threshold. Using a α quantile gives a rule with that specificity. A marker transformed to a normal distribution has $\mu + z_\alpha \sqrt{V}$ as its threshold. However, this single reference range is inefficient for any marker that behaves heterogeneously in the population (i.e., for any marker having $B_1 > 0$). Heterogeneity means that natural, or normal, marker concentrations vary between individuals, and so an implication of heterogeneity is that even though the single threshold rule gives the *population* a specificity α , subject specific specificity can vary; women with high (low) natural levels have specificity that is lower (higher) than α .

The PEB rule accounts for heterogeneity by approximating and attempts to give each woman the same selected level of specificity. This is done by using PEB statistical methods to estimate each person's normal marker

concentration, then determine the threshold defined by deviations from his/her normal levels, rather than population wide deviations. Without showing the details, it can be shown ⁶ that the PEB rule:

- (i) always has higher sensitivity than the single threshold rule with the same specificity
- (ii) has the same overall population wide specificity as the single threshold rule.

PEB refers to class of statistical procedures for estimating a group of individual means when the individuals are drawn from a common population ²⁰. Introductory descriptions of PEB are given by Casella ²¹ and Efron and Morris ²². The PEB estimator of a subject's normal marker mean is given by $\mu + (\bar{y} - \mu)B_n$, where \bar{y} is the average of her historical screens. Note that the PEB estimator is part of the threshold expression above, and that the PEB estimate is also a function of the shrinkage factor, $B_n = \tau^2 / (\sigma^2 / n + \tau^2)$. The shrinkage factor ranges from zero (when $n=0$) to one (when n is large) and the PEB estimator concomitantly ranges from population mean μ to the individual's sample average \bar{y} . The PEB estimate comes into closer agreement with the individual history when history is plentiful, but with small n the PEB estimator is closer to the population average. Thus, the PEB estimator allows the individual history to carry a greater voice when that subject has a substantial history, but anticipates regression to the mean, for subjects with little screening history.

The PEB estimator has two properties that make it useful for screening: it (I) is unbiased for estimating a person's individual mean level (II) is more precise (has a smaller variance) than the usual estimator does \bar{y} . In particular, the variance of the PEB estimator is given by $\frac{\sigma^2}{n} B_n$, which is always less than σ^2/n , the variance of \bar{y} . The PEB estimator has added precision compared to the sample average because it uses information shared among the group. This is often called "borrowing of strength" in meta-analysis. Combining these two properties with the variability of a new observation (σ^2), the threshold expression above follows. We note here that generating screening rules using the sample average, \bar{y} , instead of the PEB estimator can lead to screening rules

that do worse than the single threshold rule, and can never do better than the PEB screening rule (see McIntosh & Urban⁶ for derivation and further discussion).

Two special cases of the PEB rule are when there is no screening history, $n = 0$ (and so $B_n = 0$), and when screening history is plentiful, $n = \infty$ (and so $B_n = 1$). With the former the PEB rule and the single threshold rule coincide, and with the latter the PEB threshold becomes $\bar{y} + z_\alpha \sqrt{\sigma^2}$. Note that the latter gives a person a threshold that looks like the single threshold rule except that it uses only the *within* woman variance, as if the threshold was determined by person-specific information only. The PEB rule can therefore detect marker elevations that are $\sqrt{1-B} \times 100\%$ the size that the single threshold rule needs ($\sqrt{1-B} = \sqrt{\frac{\sigma^2}{V}}$, the ratio of deviation in PEB rule and the single threshold rule), and because the PEB rule can detect smaller elevations, it will have a higher sensitivity⁶.

In between the two extremes, when screening history n is small to modest, the PEB rule is a compromise between the population wide rule and the limiting individual rule. However, Figure 1 shows that the performance of the limiting rule does not require a large screening history and large benefits can follow after only a small number of screens. In Figure 1(a) the second screening event has a dramatic adjustment to a woman's threshold, and each subsequent screen providing comparatively smaller gains. Indeed the screening rules at the fourth screens are practically equivalent to the limiting rule for the value of B used in our example.

The reason why the PEB rule is able to increase sensitivity without changing the population specificity can be seen by examining Figure 1(a) more closely. Figure 1(a) shows that the PEB thresholds (the lines) diverge from the single threshold rule (horizontal line), depending on the mean screening history; For our example, women with a mean screening history \bar{y} under 2.6, approximately, are given lower thresholds than the single threshold rule and other women have their levels elevated from the single threshold rule. However, the average women

has $\bar{y} = 2.06$, far under this level, and so the majority of women will have their threshold lowered if the PEB rule is used. Indeed at the second screen about 84% of all women will have a lower threshold, and on subsequent screen that fraction increases to 88%, 89% and 95%, $n = 2, 3, \infty$, respectively. That is, over 95% of all women can have their cancer detected when at lower levels than a single threshold rule can. The reduction in the threshold can be dramatic, with average women receiving a cutoff of approximately 17 on the raw scale, nearly half the single threshold cutoff. The exact fraction that gets a lower threshold will be marker dependent, and is determined by the intra-class correlation B_1 ; higher values imply more women with lower thresholds. However, for any value of $B_1 > 0$, more than half the subjects will be given lower thresholds.

Of course, Figure 1 shows too that a small minority of women have higher thresholds with the PEB rule, and at first it may appear that the PEB rule may not benefit, and may even harm, these women. Further consideration shows this not to be the case, and that these women benefit as well, but in a different way. The PEB rule works by controlling the specificity of the screen for *each woman*. Compare this to the single threshold rule, which tries to control only the *population* specificity. Overall the PEB and single threshold rules will produce the same rate of false positives, but with the PEB rule each woman will share this burden equally. For ovarian cancer screening with the CA125 single threshold rule it is only a small fraction of all women, who are “unlucky” enough to have naturally high CA125 concentrations, that experiences the false positives. In our example it can be shown that 5% of all women are responsible for over the 80% of all the false positives in the population. This burden is great enough that many women may forego screening because of it. The PEB rule automatically adjusts the thresholds so that women share a controlled chance of experiencing a false positive test, and all share the same ability to detect comparable elevations. This switching from a population to an individual perspective for screening permits a dramatic increase in sensitivity without any increased burden on false positive rates.

The PEB rule presented here requires that the markers behave normally in all healthy subjects. If no convenient mathematical transformation can achieve this, then a non-parametric transformation can be made. The simplest way to do this is to tabulate the quantiles of the observed marker values, then convert them to the corresponding quantile of a standard normal distribution. Thus, a marker at the 97.5% quantile in the test population would then be given a value of 1.96, and the median value would be converted to 0.0, etc.. This gives the required behavior but may also require larger amounts of data so that the conversion is not too sparse. If formal inclusion of covariates is needed, this will require quantile regression methods, and so will add computational complexity.

Conclusion and Summary

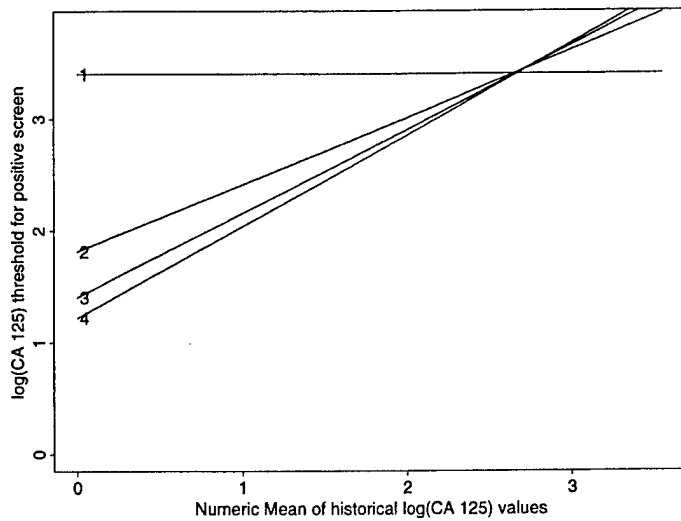
The amount of reduction gained by an individual screen depends on the marker's B , with smaller values implying smaller elevations are detectable, and more screens are needed to approach the limiting rule. This suggests strongly that a marker value of B is an important determinant of how the marker will perform when used in a cancer screening program, and thus should be reported as a routine part of marker assessment. McIntosh & Urban ⁶ give an example of a hypothetical marker having lower sensitivity than CA 125 in a single threshold rule but its larger B gives it greater sensitivity in when used longitudinally.

The data requirements of the PEB method are practical for the data availability and robust to a wide variety of marker behaviors. The specifics of the model are outlined in McIntosh & Urban ⁶, and should be consulted before applying this method to novel tumor markers. There are a few considerations that should be made before using the PEB rule in a population such as outlined in that manuscript. Specifically, further investigation should be made if the intra-class correlation B is very small, the screening interval is very short compared to the pre clinical period, and the marker grows at much less than a linear rate. If all these conditions hold, the screening rule should be modified slightly, for example by not using the PEB rule until at least 2 or 3 historical screens are available.

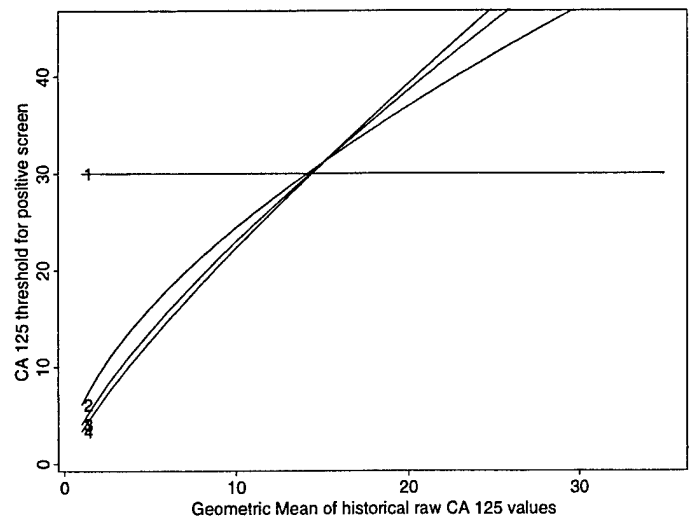
The manuscript has omitted any consideration of the sample size needed to estimate the model parameters. This was done to improve clarity. Of course in practice, larger numbers of samples will improve the precision of the estimates, but we make no recommendation regarding the number needed to evaluate the PEB Algorithm other than to state that using it to investigate the potential performance of a marker needs less data than when implementing it in an actual population. When beginning a screening program we recommend initially evaluating the model parameters on stored sera, then continually evaluating the behavior of the markers in healthy subjects until a substantial number have accumulated to be confident that it can proceed without monitoring. As new data come available the PEB algorithm (the transformation and the intra class correlation estimates) may be updated as needed. Because the PEB rule needs only healthy subjects, limits on data availability will not take long to overcome and can allow more rapid development of new tumor marker discoveries.

Acknowledgments

All authors are supported by the National Cancer Institute grant NCI 1 P50 CA83636, and Dr. Karlan is also supported by Research for Women's Cancers, Cedars-Sinai Medical Center, Los Angeles, CA



(a) A PEB screening rule on log(CA 125) scale



(b) A PEB screening rule on raw CA 125

Figure 1 Representation of the PEB screening rule for CA 125 on the log scale (left) and raw scale (right) for a specificity of approximately 0.99. Note that the PEB screening rule shown above is for a particular population, and the screening rule for another population may differ slightly, depending on the assay used, and other characteristics of the women in the study. Also note that the graph on the right uses the geometric average to determine the next threshold, not the numeric mean. Because geometric averages are lower than the numeric means, the thresholds shown in Figure 1(b) appear lower than they actually are.

List of Tables and Figures

Table 1 Tabular representation of a PEB rule calibrated to 98% specificity for a population of pre menopausal high risk women. At the initial screen, the threshold is 35 U/ml. For subsequent screens find, in the left column, the geometric average of a woman's past CA 125 levels. Read across the table to the column representing the number of historical screens available. Note that the numbers in parentheses represents approximate value of the numerical raw CA 125 levels for a corresponding geometric mean. For example, a woman with a typical CA 125 concentration between 9.3 and 10.47 may typically have a geometric mean between 8 and 9. This conversion is only approximate, and so screening should be done using the geometric mean of CA 125 only.

<i>Geometric Mean (approximate numerical mean) of historic CA 125 levels.</i>	<i>One historical screen</i>	<i>Two historical screens.</i>	<i>Three historical screens.</i>
5 (5.82)	16.10	13.65	12.64
6 (6.98)	17.96	15.56	14.67
7(8.15)	19.70	17.56	16.64
8(9.31)	21.34	19.41	18.56
9(10.47)	22.90	21.21	20.44
10(11.64)	24.40	22.95	22.28
11(12.80)	25.83	24.65	24.09
12(13.97)	27.22	26.31	25.87
13(15.13)	28.56	27.94	27.62
14(16.29)	29.86	29.54	29.34
15(17.46)	31.12	31.11	31.05
16(18.62)	32.35	32.65	32.73
17(19.78)	33.54	34.17	34.40
18(20.95)	34.72	35.67	36.04
19(22.11)	35.86	37.14	37.67

Table 2 Summary of screening program performance when using the PEB algorithm on a population of 1 million US women screened annually from age 50 through age 80 using a marker similar to CA 125 but with different amounts of marker heterogeneity. The values of B_1 reported in the literature for CA 125 fall between those in the final two columns.

Performance Measure	$B_1 = \frac{\tau^2}{\sigma^2 + \tau^2}$					
	0*	0.1	0.3	0.5	0.7	0.9
Percent of clinical cases found by screening	0.46	0.46	0.47	0.47	0.48	0.50
Mean lead time effect among all cases (years)	0.80	0.80	0.82	0.85	0.89	0.93
Fraction of cases finding a stage shift	0.35	0.35	0.36	0.37	0.40	0.42
Mortality reduction (absolute reduction from 19% 77%)	19%	19%	20%	21%	24%	26%
Mean years of life saved per case (in years)	2.53	2.55	2.67	2.87	3.13	3.44

* Equals the ST rule

Bibliography

1. Slate EH, Cronin KA. Changepoint modeling of longitudinal PSA as a biomarker for prostate cancer,. In: C. Gatsonis JSH, R.E. Kass, R. McCulloch, P. Rossi and N.D. Singpurwall, ed. Case Studies in Bayesian Statistics III, vol. III. New York: Springer Verlag, 1997; 444--56.
2. Slate E, Clark L. An application of Bayesian retrospective and prospective changepoint identification. In: C. Gatsonis BC, A. Carriquiry, A. Gelman, R Kass, I. Verdinelli and M. West., ed. Case Studies in Bayesian Statistics IV, vol. IV. New York: Springer Verlag, 1999; 511--34.
3. Morrell CH, Pearson JD, Carter H, Brant L. Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer. *Journal of the American Statistical Association* 1995;90:45--53.
4. Skates CJ, Pauler DK. Screening Based on the Risk of Cancer Calculation from Bayesian Hierarchical Change-point Models of Longitudinal Markers. *Journal of the American Statistical Association* 2001;(In Press).
5. Jacobs I, Skates S, MacDonald N, et al. Screening for ovarian cancer: a pilot randomised controlled trial. *Lancet* 1999;353(9160):1207--10.
6. McIntosh M, Urban N, A Parametric Empirical Bayes Method for Cancer Screening using Longitudinal Observations of a Tumor Marker (submitted to Biostatistics). University of Washington Department of Biostatistics, 2001.
7. Baker SG. Identifying Combinations of Cancer Markers for Further Study as Triggers of Early Intervention. *Biometrics* 2000;in press.
8. Crump K, C., McIntosh MW, Urban N, Anderson G, Karlan BY. Ovarian Cancer Tumor Marker Behavior in Asymptomatic Healthy Women: Implications for Screening. *Cancer Epidemiology, Biomarkers, and Prevention* 2000;9(10):1107-11.
9. Pitot HCC. Fundamentals of Oncology. New York, New York: Marcel Dekker, 1989.
10. Skates SJ, Singer DE. Quantifying the potential benefit of CA 125 screening for ovarian cancer. *Journal of Clinical Epidemiology* 1991;44(4-5):365-80.

11. Skates SJ, Xu FJ, Yu YH, et al. Toward an optimal algorithm for ovarian cancer screening with longitudinal tumor markers. *Cancer* 1995;76(10 Suppl):2004-10.
12. Pauler DK, Menon U, McIntosh MW, Symecko HL, Skates SJ. Factors Influencing Serum CA 125II Levels in Healthy Postmenopausal Women. *Accepted to Cancer Epidemiology, Biomarkers and Prevention* 2001.
13. Davelaar EM, van Kamp GJ, Verstraeten RA, Kenemans P. Comparison of seven immunoassays for the quantification of CA125 antigen in serum. *Clinical Chem.* 1998;44:1417-22.
14. Karlan BY, Raffel LJ, Crvenkovic G, et al. A multidisciplinary approach to the early detection of ovarian carcinoma: rationale, protocol design, and early results. *American Journal of Obstetrics & Gynecology* 1993;169(3):494-501.
15. Urban N, McIntosh M, Clarke L, et al. Socioeconomics of ovarian cancer. In: Jacobs I, ed. *Ovarian Cancer*: Oxford University Press, 2001.
16. Urban N, Drescher C, Clarke L, Kiviat N. Cost-Effective Analysis of Ovarian Cancer Screening Strategies. *Hungarian Journal of Gynecologic Oncology (in Hungarian)* 1997;2:169-80.
17. Urban N, Drescher C, Etzioni R, Colby C. Use of a stochastic simulation model to identify an efficient protocol for ovarian cancer screening. *Controlled Clinical Trials* 1997;18:251-70.
18. DePriest PD. Ovarian Cancer Screening in Asymptomatic Post-Menopausal women. *Gynecol Oncol* 1993;51:7-11.
19. DePriest PD, Gallion HH, Pavlik EJ, Kryscio RJ, van Nagell JR, Jr. Transvaginal sonography as a screening method for the detection of early ovarian cancer. *Gynecologic Oncology* 1997;65(3):408-14.
20. Morris CN. Parametric Empirical Bayes inference: Theory and applications (with discussion). *Journal of the American Statistical Association* 1983;78(381):47-65.
21. Casella G. An introduction to empirical Bayesian data analysis. *American Statistician* 1985;39:83-7.
22. Efron B, Morris CN. Stein's paradox in statistics. *Scientific American* 1977;236(5):119-27.

Appendix H

Enrollment Forms

- ORCHID Brochure
- ORCHID Consent Form
- ORCHID Medical Records Release
- ORCHID Questionnaire

Who conducts the study?

This work is a collaboration among experts in oncology, immunology, molecular biology, and statistical methods representing Fred Hutchinson Cancer Research Center, the University of Washington, Virginia Mason Research Center and Swedish Medical Center. Nicole Urban, ScD, of the Fred Hutchinson Cancer Research Center, is the Principal Investigator of the study. Funding for this project is awarded by the Department of Defense Ovarian Cancer Research Program to the Fred Hutchinson Cancer Research Center.

What is the Marsha Rivkin Center for Ovarian Cancer Research™?

In September 1989, Marsha Rivkin was diagnosed with ovarian cancer. Four years later, at the age of 49, she passed away leaving behind five daughters and her husband of 29 years, Dr. Saul Rivkin, an oncologist at Swedish Medical Center in Seattle. In Marsha's honor, and in recognition of the importance of continued work against this deadly disease, the Rivkin family founded the Marsha Rivkin Center for Ovarian Cancer Research.

The mission of the Marsha Rivkin Center is to improve the outcomes for women diagnosed with ovarian cancer, and those at risk for the disease, by encouraging collaborative scientific research, increased education and awareness, and community participation. The Swedish Medical Center Foundation raises funds for the Marsha Rivkin Center, some of which were used to support costs for the pilot phases of the ORCHID study.

What is the Fred Hutchinson Cancer Research Center?

The Fred Hutchinson Cancer Research Center is an independent, non-profit research institution dedicated to developing new knowledge to eliminate cancer. The Hutchinson Center is designated by the National Cancer Institute as a comprehensive cancer center. The Hutchinson Center is a world leader in laboratory, treatment and prevention research.

How can I get more information?

Contact the study office at **(206) 215-6200** or write to the address below:

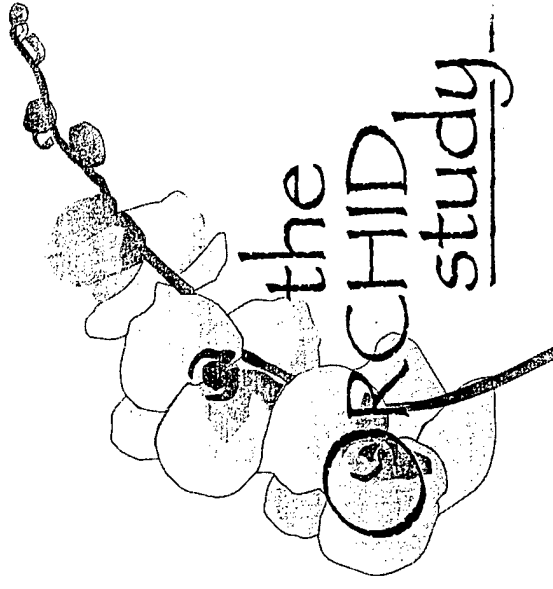
ORCHID Study
Marsha Rivkin Center
for Ovarian Cancer Research
1221 Madison Street, Suite 1410
Seattle, WA 98104



SWEDISH MEDICAL CENTER

VIRGINIA MASON

Research Center



Improving Ovarian
Cancer Screening

What is the ORCHID Study?

ORCHID stands for Ovarian Research Collaboration Helping to Improve Detection. The goal of the ORCHID study is to improve women's health by developing new screening tests to find ovarian cancer early, when it is easiest to treat.

Who may participate in the study?

You are eligible to participate if you:

- Are at least 18 years of age
- Have one or both ovaries
- Are willing to provide blood and/or ovarian tissue specimens

Women with ovarian cancer, women with benign ovarian disease (not cancer), and women with disease-free ovaries are invited to participate in this research.

What does participation in ORCHID involve?

Participation in ORCHID involves filling out a 15-minute questionnaire, giving permission for us to look at your medical records, and donating blood and/or tissue to be used in research analysis. Only women undergoing ovarian-related surgery at a participating hospital are asked to donate tissue. Scientists will look for molecular changes in the tissue and blood of women with and without ovarian cancer to aid in the development of a screening test for ovarian cancer.

What is ovarian cancer screening?

Ovarian cancer screening consists of medical tests that help find ovarian cancer in an early stage when there are no clinical signs of disease. Scientists with the ORCHID study are working to develop accurate methods of screening that can be used in the general population.

If I choose to participate in the study, how will I donate blood?

You can give blood during your regular office visit to your doctor. Women who are undergoing surgery at a participating hospital can donate blood at the time of their surgery.

If I choose to participate in the study, how will I donate tissue?

Some women require surgery to determine the best way to treat their medical condition. If you have surgery as part of your regular medical care and decide to participate in ORCHID, you may choose to donate tissue samples to the study.

During surgery, doctors remove tissue from the woman's ovary and send it to a lab for evaluation. The doctor in the lab, a pathologist, examines a portion of the tissue to help determine the best course of care. It is often not necessary to examine all of the tissue removed, so the remaining unexamined tissue is generally disposed of appropriately. If the patient is a participant in ORCHID and chooses to donate tissue, we will keep a portion of the tissue that is normally discarded and use it in our research.

How will participating in ORCHID affect my medical care?

Participating in this study will not affect your medical care. You will receive the same excellent medical care whether you participate in this study or not.

What will I get out of the study?

Participation in this study will not provide any direct benefit to you, but your contribution to this research effort may help other women at risk for ovarian cancer in the future.

Will I have to pay to participate? Will there be costs to me?

Participation in this study will not cost you any money. We will pay for the costs of blood and tissue collection. Neither you nor your insurance company will be billed for any procedures related to the collection of blood and tissue samples for the ORCHID study.

How much time will I have to give to the study?

The time required to participate in this study will be approximately 15-20 minutes. We will ask you to read and sign a consent form, complete a questionnaire, and discuss your questions with a member of the research team. Blood and tissue donations generally occur during a regular office visit or medical procedure and do not require extra time.

CONSENT TO PARTICIPATE IN RESEARCH STUDY:

ORCHID - Ovarian Research Collaboration Helping to Improve Detection
“Use of Novel Technologies to Identify and Investigate Molecular Markers for Ovarian
Cancer Screening and Prevention”

Conducted by Investigators at the Fred Hutchinson Cancer Research Center (FHCRC), University of Washington (UW) and Virginia Mason Research Center (VM). Funding by the Marsha Rivkin Center for Ovarian Cancer Research and the United States Army Medical Research and Materiel Command (USAMRMC).

NAME	POSITION	INSTITUTION
Nicole Urban, ScD	Principal Investigator	FHCRC
Garnet Anderson, PhD	Co-Investigator	UW
Charles W. Drescher, MD	Co-Investigator	FHCRC
Leona Holmberg, MD	Co-Investigator	FHCRC
Leroy Hood, PhD, MD	Co-Investigator	UW
Nancy Kiviat, MD	Co-Investigator	UW
Jane Kuypers, PhD	Co-Investigator	UW
Brad Nelson, PhD	Co-Investigator	VM
Mary Anne Rossing, PhD	Co-Investigator	FHCRC
Michel Schummer, PhD	Co-Investigator	UW

Location(s) of Study:**Fred Hutchinson Cancer Research Center (FHCRC)**

1100 Fairview Avenue North
 Seattle, WA 98109
 (206) 667-5000

Marsha Rivkin Center for Ovarian Cancer Research

1221 Madison Street, Suite 1410
 Seattle, WA 98104
 (206) 386-2419

Pacific Gynecology Specialists

1101 Madison Street, #1500
 Seattle, WA 98104
 (206) 215-3200

University of Washington (UW)

School of Medicine
 Seattle, WA 98195
 (206) 543-2100

Virginia Mason Research Center (VM)

1000 Seneca Street
 Seattle, WA 98101

Providence Seattle Medical Center

500 17th Avenue
 Seattle, WA 98124
 (206) 320-2000

PURPOSE AND BENEFITS

You have been asked to participate in a study of ovarian tissue. This study is a collaborative effort of several investigators. The goal of this research study is to establish a repository of gynecologic specimens that can be used to develop new screening tests to detect ovarian cancer before it spreads outside the ovary and becomes very difficult to cure. Investigators (research doctors) will study tumor tissue and blood from women with ovarian cancer, women with benign ovarian conditions, and women with no ovarian disease in order to identify genetic changes, or other changes that can be found in the blood, that are associated only with ovarian cancer. Results of this research will be used to develop methods for early diagnosis and prevention of ovarian cancer or other gynecologic cancers, and/or to provide better treatment for ovarian cancer. In addition, it will help scientists to better understand the biology of ovarian tumors. There is no direct benefit to you for participating in this study.

We are asking all women who undergoing ovarian related surgery to consider making their tissue and/or blood specimens available to investigators to carry out this research, which includes analysis of the cells at the molecular level (i.e., the smallest particles or components of the cell). This study is designed to look at differences between normal cells and the genetically altered cells that may give rise to cancer. These genetic factors may be inherited or they may be changes that result from environmental exposures. If you agree to participate in this study, a small amount of your white blood cells may be "grown" in the laboratory to develop a "cell line" that can be used for research.

PROCEDURES

You can participate in one of two ways in this study: by providing blood and tissue specimens, or by providing blood specimens only. **This consent form applies only to patients who wish to provide blood & tissue specimens to the research study (OPTION A).**

If you are scheduled to undergo ovarian related surgery at Swedish Medical Center or Providence Seattle Medical Center, you will be asked to provide ovarian tissue and blood samples. Acceptance or rejection of your participation in this study will in no way affect the treatment you receive for your condition, nor will it affect the outcome of your treatment. If you agree to participate, your surgery will be conducted exactly as if you were not going to participate.

In order to obtain the best results possible, we will ask you to fill out a medical questionnaire. The data collected in this questionnaire will focus on known or suspected risk factors for ovarian cancer, and will include questions on menstrual and reproductive history, birth control history, family and personal history of ovarian, breast and other cancers and sociodemographic factors. We will also need to review your medical chart for additional health information. The information provided for the questionnaire and gained from the medical chart will only be identifiable by a special study number. Your name will not appear on any reports of the results of these studies.

We will ask you to provide a sample of blood (approximately 40 ml or 3-4 tablespoons). Whenever possible, the blood draw will take place at the time of your surgery, and will be conducted by your anesthesiologist. You may also choose to have blood drawn for this study by a certified phlebotomist at the time of your pre-operative visit, or in the pre-operative holding area on the day of surgery. Blood provided to this study will be obtained once only.

During surgery, your surgeon will first remove the ovarian tissue. A certain amount of tissue specimen is needed to assist your surgeon in making a diagnosis. If there is not enough tissue specimen to both make a diagnosis and contribute to the research study, all of the tissue removed will be used to make a diagnosis. If there is enough tissue specimen to both make a diagnosis AND contribute to the research study, tissue and blood samples will be provided to the study tissue collection specialist.

Results of the analysis of your tissue are of no known value with regard to your personal health and will not be made known to you or your treating physician. The biological specimens you provide to this study will be stored at the designated study repository in Seattle, Washington for an indefinite period of time. The specimens will then be requested by laboratory investigators for molecular analysis. No specific genetic information will be generated that could be applied to your medical care, and therefore you will be given no genetic information after the studies are complete. These tests are experimental and are of unknown value in the diagnosis and treatment of ovarian cancer.

RISKS, STRESS OR DISCOMFORT

Taking blood may cause temporary discomfort, and bruising may form at the site of needle entry. Any tissue that is obtained for the purposes of this research study is collected only after it has been removed for the purposes of the surgical procedure that you are receiving. The molecular and genetic analyses that will be conducted on tissue specimens may discover known or unknown genetic alterations. The results of these analyses will not be released to you or your personal care provider. These tests are experimental and are of unknown value in the diagnosis and treatment of ovarian cancer. However, this non-identified information may be shared with investigators at other, additional, institutions. Not knowing these results may be of potential discomfort to you or may not be of concern to you. Being part of the study should not cause any additional stress for you around the time of your operation. The duration of your participation will vary depending on how much time is needed to 1) answer questions about the research to your satisfaction; 2) complete the enrollment forms and 3) draw your blood.

ALTERNATIVES

The alternative to participating in this study is not to provide biological specimens (tissue and blood) to the investigators in this research study.

USE OF SPECIMENS

We may store your blood and tissue samples to use in future research. If we want to use them for a research purpose not described in this consent form, we will send our request to the Institutional Review Board. This Board protects the rights and welfare of research subjects like you. The Board will determine if we need to contact you and ask your consent to do the research. Future research using your blood and tissue could lead to the development of commercial products. You will not share in any profits that this work may produce.

Any specimens requested for use in future research will be subject to review and approval by the Investigators of this study and the FHCRC Institutional Review Board. As such, this study has established a three-tier review process to ensure that patient's rights are protected should specimens ever be requested by other research studies or commercial entities. Commercial entities often conduct their own research studies as well as provide funding for research studies. If specimens are requested from commercial entities, or research studies in collaboration with a commercial entities, approval must also be obtained from the FHCRC Human Specimens Committee. Under no circumstances will patient identifying information be made available to future research studies.

COSTS AND COMPENSATION

It should be understood that whether or not you participate, all medical expenses relating to your surgical procedures will be paid by you and/or your insurance company. There are no additional costs to you for participating in this study.

The Department of Defense (USAMRMC) is funding this research project. Should you be injured as a direct result of participating in this research project, you will be provided medical care, at no cost to you, for that injury. You will not receive any injury compensation, only medical care. You should also understand that this is not a waiver or release of your legal rights.

You should discuss this issue thoroughly with the Principal Investigator before you enroll in this study.

There is no financial compensation for participation in this program. By agreeing to participate in this research study, you understand that you waive any claim to monetary gain or financial benefit as relates to this study. If you have any questions regarding your costs, financial responsibilities, and/or medical insurance coverage for this activity, please contact your primary care physician or Pacific Gynecology Specialists at (206) 215-3200.

OTHER INFORMATION

We have taken extensive precautions to maintain the confidentiality of all study records. All records containing personal information will be kept confidential as provided by law. Strict protocols will be followed to maintain the confidentiality of any identified patient information. Study records will be maintained indefinitely for the purpose of analysis and follow up. Your personal identity will not be revealed in any publication or release of results. Please be aware that representatives from the U.S. Army Medical Research and Materiel Command will also have access to the study records, and may inspect the records of the research in their duty to protect human subjects in research.

Your participation in this study is *voluntary*. Once enrolled, you may discontinue participation at any time for any reason, without notice. Declining to participate in this study will involve no penalty or loss of benefits to which you may otherwise be entitled. You will continue to receive your usual medical care even if you decide not to participate in this study. Questions about this study may be addressed to the Principal Investigator, your surgeon, or research staff. Your participation is critical to the success of this study. We want you to know that you are the most important part of this study. Without your participation, this type of study will not be possible. We have tried to make participation as convenient and easy as possible for you. Please let us know if there are ways you think it would be easier for you and others to participate in this study.

If you have any questions about the research study, or in the event of a research-related injury, please contact the Principal Investigator, Nicole Urban in the Division of Public Health Sciences at Fred Hutchinson Cancer Research Center at (206) 667-4677. If you have any questions specifically about your rights as a research participant, please contact Karen Hansen in the Institutional Review Office of the Fred Hutchinson Cancer Research Center at (206) 667-4867.

Investigator's Statement

I have furnished a qualified and trained study nurse or research staff member to provide an explanation of the above research program. The patient was given an opportunity to discuss the procedures, including possible alternatives, with this person or their physician and to ask any additional questions. In addition the patient was provided with my name and telephone number and informed that she could contact me for any additional questions. A signed copy of the consent form has been given to the patient.

Nicole Urban, ScD
Principal Investigator

Date

Patient's Statement:

I agree to this study and to the conditions outlined in this consent form. I have had the opportunity to ask questions about the study and my participation and about the need for access to my medical records. They have been answered to my satisfaction. I understand future questions I may have about the research will be answered by one of the investigators listed above and that any questions I have about my rights as a research participant will be answered by the person identified above. I give permission for my medical records to be available for review and copying to the appropriate physicians and personnel for this study at the Fred Hutchinson Cancer Research Center, Pacific Gynecology Specialists, University of Washington and United States Army Medical Research and Materiel Command. I understand that there is a possibility that the blood, tissue and bodily fluids (specimens) which I am providing under this study may also be used in other research studies and could potentially have some commercial applicability.

 Patient's Name (printed)

 Patient's Signature

 Date

 Patient's Permanent Address:

 City,

 State

 Zip

 Witness Name (printed)

 Witness Signature

 Date
Patient's Statement of Donation:

I voluntarily and freely donate any and all blood, tissues and bodily fluid to the Fred Hutchinson Cancer Research Center and hereby relinquish all right, title, and interest to said items.

 Patient's Name (printed)

 Patient's Signature

 Date

 Witness Name (printed)

 Witness Signature

 Date

Copies to:

 Patient
 Medical Records
 Research File

Participant Enrollment Form & Medical Records Release

Home Phone: () -

Med. Rec. ID: _____ Physician ID: _____ Enrolled by: _____

ORCHID

How to Fill Out this Questionnaire

This questionnaire asks you about your general health, medical history, family history, and some lifestyle habits. It takes about 20 minutes to complete. All of your answers will be kept strictly confidential.

Before you begin, please note the following.

1. If a question asks you to “check one” answer, please check the one that best describes you.
2. If you are unsure about how to answer a question, make your best guess. If you cannot provide a guess or estimate, please check or write “Don’t know”.
3. Some questions ask about “full” and “half” brothers and sisters. A “full” brother or sister has the same two parents as you. A “half” brother or sister has only one of the same parents as you.
4. You may refuse to answer any of the questions in this questionnaire and can stop at any time.

Thank you for taking the time to complete this questionnaire. Your participation is greatly appreciated!

QUESTIONS 1-6

1. What is today's date?

Month

Day

Year

2. What is your date of birth?

Month

Day

Year

3. In what country were you born?

4. How tall are you?

Feet

Inches

5. How much did you weigh one year ago?

Pounds

6. What is your marital status now?

☐₁ Married or living as married

☐₂ Widowed

☐₃ Divorced

☐₄ Separated

☐₅ Never Married

QUESTIONS 7-23 ARE ABOUT YOUR REPRODUCTIVE AND MEDICAL HISTORY.

7. How old were you when you had your first menstrual period?

Years old

8. During most of your life, were your periods regular; that is, did they occur about once a month? **(Do not include any times when you were pregnant or taking birth control pills.)**

☐ ₁ No

☐ ₂ Yes

☐ ₃ Sometimes regular, sometimes irregular



8.1 How old were you when your periods first became regular?
(Your best guess.)

Years old

9. Between the time you had your first period and your last period, did you ever go without any periods for at least one year? **(Do not count times when you were pregnant or breastfeeding.)**

☐ ₁ No

☐ ₂ Yes



9.1 Between your first menstrual period and your last, all together, about how long did you go without having your period?
(Again, do not count times when you were pregnant or breastfeeding.)

Months

Go to the next page.

10. Have you ever been pregnant?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 11)

10.1 Including live births, stillbirths, miscarriages, abortions, tubal, or ectopic pregnancies, how many times have you been pregnant? **(If you are currently pregnant, be sure to count this pregnancy.)**

Number of pregnancies

10.2 How many of your pregnancies lasted 6 or more months?

Number of pregnancies

10.3 How many live births resulted from these pregnancies?

Number of live births

10.4 How old were you when you had your first live birth or stillbirth?

Years old

11. Have you ever tried to become pregnant for more than 1 year without becoming pregnant?

☐₁ No

☐₂ Yes

☐₈ Don't know

11.1 Did you visit a doctor or clinic because you didn't get pregnant?

☐₁ No

☐₂ Yes

11.2 Was a reason found for why you did not become pregnant?

☐₁ No

☐₂ Yes

☐₈ Don't know

11.3 What was the reason you did not become pregnant?
(Mark all that apply)

☐₁ Problem with your hormones or ovulation
(Producing eggs)

☐₂ Problem with your tubes or uterus

☐₃ Endometriosis

☐₄ Other problem with you (Specify): _____

☐₅ Problem with partner

☐₈ Don't know

Go to the next page.

12. Did you ever breastfeed or nurse any children for at least one month?

☐₁ No

☐₂ Yes

12.1 How many children did you breastfeed?	<input type="text"/>	Number of children
12.2 How old were you when you <u>first</u> breastfed a child?	<input type="text"/>	Years old
12.3 How old were you when you <u>last</u> breastfed a child?	<input type="text"/>	Years old
12.4 Thinking about all the children you breastfed, how many months <u>total</u> did you breastfeed? (Your best guess)	<input type="text"/>	Number of months

13. Have you ever used birth control pills for any reason?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 14)

13.1 How old were you when you first used birth control pills?	<input type="text"/>	Years old
13.2 How many years altogether have you used birth control pills?	<input type="text"/>	Number of years

14. Did you ever have an operation to have your tubes tied to prevent pregnancy?

☐₁ No

☐₂ Yes

14.1 How old were you when you had your tubes tied?
<input type="text"/> Years old

Go to the next page.

15. Have you had a menstrual period in the last 12 months?

☐₁ No

☐₂ Yes

☐₈ Don't know

16. When was your last menstrual period (best guess)?

Month

Year

17. How would you describe your current menstrual periods? Mark the one statement that best describes your situation.

☐₁ Still having periods, or currently pregnant or nursing

☐₂ Possibly going through menopause (the change of life)

☐₃ Periods stopped by themselves (natural menopause)

☐₄ Periods stopped by surgery (removal of uterus, ovaries, or both)

☐₅ Still having periods due to hormone replacement therapy

☐₈ Other (Specify): _____

18. Have you had a hysterectomy (surgical removal of your uterus)?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 19)



18.1 How old were you when you had a hysterectomy?

Years old

19. Have you ever used estrogen pills, creams, or skin patches (for example, Premarin, Estrace, Ogen, Estraderm)? Sometimes estrogens are given to treat symptoms of menopause (e.g. hot flashes, night sweats), to prevent osteoporosis (thin or brittle bones), or to prevent heart disease. (Include all hormones except pills used for birth control.)

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 20)



19.1 How old were you when you first used estrogen?

Years old

19.2 How many years altogether have you used estrogen?

Number of years

19.3 When did you last use estrogen?

Month

Year

Go to the next page.

20. Did a doctor ever say that you had any of the following conditions?

20.1	Diabetes, high blood sugar, or sugar diabetes	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.2	Inflammatory bowel disease, colitis, or Crohn's disease	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.3	Chronic lung disease, bronchitis, or emphysema	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.4	Heart failure or congestive heart failure	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.5	Heart attack, coronary, or myocardial infarction	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.6	High blood cholesterol requiring pills	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.7	High blood pressure (hypertension)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.8	Stroke or brain hemorrhage	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.9	Liver disease or cirrhosis	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.10	Chronic kidney disease or kidney failure	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.11	Depression or anxiety requiring pills	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.12	Osteoporosis (weak, thin, or brittle bones)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.13	Asthma	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.14	Thyroid problem (not cancer)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.15	Fibroids (benign tumors) in your uterus or womb	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.16	Endometriosis	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.17	Benign breast disease or fibrocystic breast disease	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.18	Pelvic inflammatory disease, PID or pelvic infection	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.19	Rheumatoid arthritis, SLE or scleroderma	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.20	Polycystic ovarian disease, PCO or sclerocytic ovaries	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
20.21	Ovarian cyst	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know

21. Did a doctor ever say that you had breast cancer?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 22)

21.1 How old were you when you were first diagnosed with breast cancer?

Years old

21.2 Was the breast cancer found in one or both breasts?

One

☐₁

Both

☐₂

Don't know

☐₈

22. Did a doctor ever say that you had ovarian cancer?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 23)

22.1 How old were you when you were first diagnosed with ovarian cancer?

Years old

23. Did a doctor ever say that you had any other type of cancer?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 24)

23.1 What type of cancer did you have?

How old were you when you were first diagnosed with this cancer?

Years old

Years old

Years old

Go to the next page.

QUESTIONS 24-33 ARE ABOUT YOUR FAMILY.

24. Are you adopted?

☐₁ No

☐₂ Yes

☐₈ Don't know

25. Are you a twin?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 26)

25.1 Are you an identical or fraternal twin?

Identical

☐₁

Fraternal

☐₂

Don't know

☐₈

Go to the next page.

26. Please complete the following questions on **female** family members related to you by blood (that is, not related to you by marriage or by adoption). ***Include all blood relatives, both those living and deceased.*** For each relative type, write the number in your family. (Write “00” in the box if the answer is “none”, and “88” if the answer is “Don’t know”).

26.1 How many full sisters do you have?	<input type="text"/>	Full sisters
26.2 How many half sisters do you have?	<input type="text"/>	Half sisters
26.3 How many daughters do you have, not including step daughters?	<input type="text"/>	Daughters
26.4 How many full sisters does your mother have?	<input type="text"/>	Aunts
26.5 How many half sisters does your mother have?	<input type="text"/>	Aunts
26.6 How many full sisters does your father have?	<input type="text"/>	Aunts
26.7 How many half sisters does your father have?	<input type="text"/>	Aunts
26.8 How many nieces do you have?	<input type="text"/>	Nieces

27. Have any of the following female relatives related to you by blood ever been diagnosed with breast cancer?

27.1 Mother	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don’t know
27.2 Grandmother(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don’t know
27.3 Full sister(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don’t know
27.4 Half sister(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don’t know
27.5 Daughter(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don’t know
27.6 Aunt(s) on father’s side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don’t know
27.7 Aunt(s) on mother’s side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don’t know
27.8 Niece(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don’t know

28. Have any of the following female relatives related to you by blood ever been diagnosed with ovarian cancer?

28.1 Mother	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
28.2 Grandmother(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
28.3 Full sister(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
28.4 Half sister(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
28.5 Daughter(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
28.6 Aunt(s) on father's side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
28.7 Aunt(s) on mother's side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
28.8 Niece(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know

29. Please complete the following questions on **male** family members related to you by blood. ***Include all blood relatives, both those living and deceased.*** For each relative type, write the number in your family. (Write "00" in the box if the answer is "none", and "88" if the answer is "Don't know".)

29.1 How many full brothers do you have?	<input type="text"/>	Full brothers
29.2 How many half brothers do you have?	<input type="text"/>	Half brothers
29.3 How many sons do you have, not including step sons?	<input type="text"/>	Sons
29.4 How many full brothers does your mother have?	<input type="text"/>	Uncles
29.5 How many half brothers does your mother have?	<input type="text"/>	Uncles
29.6 How many full brothers does your father have?	<input type="text"/>	Uncles
29.7 How many half brothers does your father have?	<input type="text"/>	Uncles
29.8 How many nephews do you have?	<input type="text"/>	Nephews

30. Have any of the following male relatives related to you by blood ever been diagnosed with prostate cancer?

30.1 Father	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
30.2 Grandfather(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
30.3 Full brother(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
30.4 Half brother(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
30.5 Son(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
30.6 Uncle(s) on father's side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
30.7 Uncle(s) on mother's side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
30.8 Nephew(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know

31. Have any of the following male relatives related to you by blood ever been diagnosed with breast cancer?

31.1 Father	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes	<input type="checkbox"/> ₈ Don't know
31.2 Grandfather(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
31.3 Full brother(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
31.4 Half brother(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
31.5 Son(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
31.6 Uncle(s) on father's side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
31.7 Uncle(s) on mother's side	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know
31.8 Nephew(s)	<input type="checkbox"/> ₁ No	<input type="checkbox"/> ₂ Yes → How many? <input type="text"/>	<input type="checkbox"/> ₈ Don't know

32. Have any other types of cancer or malignant tumors occurred in any of the family members (blood relatives) you listed in questions 26 and 29?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 33)

32.1 What other type(s) of malignant tumors? (Check all that apply and list how many relatives had each type.)

☐ Bladder

How many?

☐ Brain

How many?

☐ Colon

How many?

☐ Kidney

How many?

☐ Leukemia

How many?

☐ Lung

How many?

☐ Lymphoma

How many?

☐ Melanoma

How many?

☐ Pancreas

How many?

☐ Skin (not melanoma)

How many?

☐ Thyroid

How many?

☐ Uterus, womb, or endometrium

How many?

☐ Cervix

How many?

☐ Other cancer(s) or malignant tumors

How many?

(Specify): _____

33. Do any other types of health problems, disorders or diseases tend to run in your family (in two or more blood relatives)?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 34)

33.1 What is the problem or disease?

How many members of your family have this problem or disease?

Relatives

Relatives

Go to the next page.

QUESTIONS 34-35 ASK ABOUT YOUR LIFESTYLE HABITS.

34. Have you smoked a total of 100 cigarettes or more in your lifetime?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 35)

34.1 How old were you when you first started smoking cigarettes? : Years old

34.2 Do you smoke cigarettes now?

☐₁ No

☐₂ Yes → (Go to Question 34.4)

34.3 How old were you when you last stopped smoking cigarettes?

:
Years old

→ (Go to Question 34.5)

34.4 How many cigarettes each day do you smoke? : Per day

34.5 How many total years have you smoked (or did you smoke)? : Years

35. Did you ever drink alcoholic beverages (beer, wine, or liquor) at least once a month for 6 months or more?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 36)

35.1 Have you had any alcoholic beverages during the past six months?

☐₁ No

☐₂ Yes

☐₈ Don't know → (Go to Question 36)

35.2. During the past six months, how many alcoholic beverages did you usually have?

☐₁ None or less than one per month

☐₂ 1-3 each month

☐₃ 4-6 each month

☐₄ 1 each day

☐₅ 2-4 each day

☐₆ 5 or more each day

Go to the next page.

QUESTIONS 36-38 ASK ABOUT YOUR BACKGROUND

36. What is the highest level of education that you have completed? **(Check one)**

- ☐₁ 8th grade or less
- ☐₂ 9th - 11th grade
- ☐₃ Graduated from high school (or GED)
- ☐₄ Vocation, technical, or business training
- ☐₅ Some college or junior college
- ☐₆ Graduated from college
- ☐₇ Attended graduate or professional school

37. How would you describe your racial or ethnic group? If you are of mixed descent, with which group do you identify most? **(Check one)**

- ☐₁ Native American, American Indian, or Alaska Native
- ☐₂ Asian or Pacific Islander
- ☐₃ Black or African American (not of Hispanic origin)
- ☐₄ Hispanic/Latino (ancestry is Mexican, Cuban, Puerto Rican, Central American, or South American)
- ☐₅ White (not of Hispanic origin)
- ☐₈ Other (Specify): _____

38. What was your job status one year ago? Check the one that best describes you.

- ☐₀₁ Working full-time (35 hours per week or more)
- ☐₀₂ Working part-time
- ☐₀₃ On temporary leave
- ☐₀₄ Not working for pay
- ☐₀₅ Retired
- ☐₀₆ Disabled, unable to work
- ☐₀₇ In school, not working for pay
- ☐₀₈ Homemaker, raising children, care of others
- ☐₁₀ Other (Specify): _____

QUESTIONS 39-49 ASK ABOUT FEELINGS YOU MAY HAVE ABOUT POSSIBLE DIFFICULTIES IN YOUR LIFE. SOME OF THEM ALSO ASK ABOUT YOUR SOURCES OF SOCIAL SUPPORT FOR DEALING WITH THOSE DIFFICULTIES.

In the last month, how often have you: **(Check one box on each line.)**

	Never	Almost never	Sometimes	Fairly often	Very often
39. Felt you were unable to control important things in your life?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
40. Felt confident about your ability to handle personal problems?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
41. Felt things were going your way?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
42. Felt difficulties were piling up so high that you could not overcome them?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅

People sometimes look to others for help, friendship, or other types of support. Next are some questions about the support that you have. How often is each of the following kinds of support available to you if you need it? **(Check one box on each line.)**

	None of the time	A little of the time	Some of the time	Most of the time	All of the time
43. Someone you can count on to listen to you when you need to talk.	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
44. Someone to give you good advice about a problem.	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
45. Someone to take you to the doctor if you need it.	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
46. Someone to help with daily chores if you are sick.	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
47. Someone to share your most private worries and fears.	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
48. Someone to do something fun with.	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅
49. Someone to love you and make you feel wanted.	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	<input type="checkbox"/> ₅

LABEL

Interest in Future Research Studies

Your participation in this research study has been very important. If you would like to be contacted about participating in future research studies on cancer and its prevention, please indicate so by checking the appropriate box and completing the following information:

☐₀₁ I am interested in being contacted about future research studies.

☐₀₂ I am *not* interested in being contacted about future research studies.

Your signature: _____

Name (please print): _____
First name *Last name*

Address: _____
Street *Apt. #*

City *State* *Zip*

Telephone: (_____) _____

If you indicate an interest in participating in future studies, your name will be kept for five (5) years so that researchers may contact you when a study is initiated. Agreeing to be contacted does not mean that you have to participate in a future research study – only that you will consider it.

Thank you again for your cooperation.

*Thank you very
much for completing
this questionnaire.*

Appendix I
Specimen Related Forms

- ORCHID Specimen Collection Form
- ORCHID Blood Processing Form

ORCHID – Blood Specimen Form

UPN: _____

Patient name: _____
LAST FIRST MI

DynaCare requisition no.: _____

Blood drawn in: ☐ surgery ☐ clinic

Date of blood draw: ____ / ____ / ____

Processing & Transport

☐ No. tubes submitted for processing: ☐ Red top ☐ EDTA

☐ Transported to freezer holding area after processing _____

INITIALS

*In the spaces below, record the 6-digit number on the label of each vial received from the processing labs.
 Please note if any vial is only partially full, or if the contents are hemolyzed or lipidemic.*

Serum

**WBC
Pellets**

Plasma

--	--

Appendix J
Specimen Characterization Forms

- ORCHID Histology I
- ORCHID Histology II
- ORCHID Clinical Data Form
- POCRC Clinical Status Form

ORCHID – Specimen Histology Report I

UPN: _____

Date of analysis: ____/____/____
Form completed by: _____

Clinical diagnosis: _____

Site of Primary Ca.: ☐ Ovarian
☐ Other:

Pathology Diagnosis:

☐ Normal

☐ ₀₃ Normal ovarian/tubal tissue
☐ ₀₅ Other normal:

☐ Malignant

Serous

☐ ₁₆ Serous carcinoma of LMP
☐ ₁₇ Serous carcinoma
☐ ₁₉ Serous cystadenofibroma
☐ ₂₀ Serous adenofibroma

Mucinous

☐ ₂₈ Mucinous adenocarcinoma of LMP
☐ ₂₉ Mucinous carcinoma
☐ ₃₀ Malignant mucinous adenofibroma
☐ ₃₁ Malignant mucinous cystadenofibroma

☐ ₅₇ Other:

Endometrioid

☐ ₃₇ Endometrioid carcinoma of LMP
☐ ₃₈ Endometrioid adenocarcinoma

Other

☐ ₄₀ Malignant adenosarcoma (mesodermal)
☐ ₄₁ Mesodermal (mullerian) mixed tumor, homo.
☐ ₄₂ Mesodermal (mullerian) mixed tumor, hetero.
☐ ₄₇ Clear cell carcinoma of LMP
☐ ₄₈ Clear cell carcinoma
☐ ₅₂ Brenner tumor of LMP
☐ ₅₃ Malignant Brenner tumor
☐ ₅₄ Undifferentiated carcinoma
☐ ₅₅ Adenocarcinoma, NOS

☐ Benign

Serous

☐ ₁₁ Serous cystadenoma
☐ ₁₂ Serous adenofibroma
☐ ₁₃ Serous cystadenofibroma
☐ ₁₄ Proliferating serous adenofibroma
☐ ₁₅ Proliferating serous cystadenofibroma

Mucinous

☐ ₂₅ Mucinous cystadenoma
☐ ₂₆ Mucinous adenofibroma
☐ ₂₇ Mucinous cystadenofibroma

Non-neoplastic

☐ ₀₆ Paraovarian cyst
☐ ₀₇ Functional cyst
☐ ₀₈ Corpus luteum
☐ ₀₉ Inflammatory lesion
☐ ₁₀ Endometriosis

Endometrioid

☐ ₃₂ Endometrioid cystadenoma
☐ ₃₃ Endometrioid adenofibroma
☐ ₃₄ Endometrioid cystadenofibroma
☐ ₃₅ Proliferating endometrioid adenofibroma
☐ ₃₆ Proliferating endometrioid cystadenofibroma

Other

☐ ₃₉ Benign adenofibroma (mesodermal)
☐ ₄₃ Clear cell adenofibroma
☐ ₄₄ Clear cell cystadenofibroma
☐ ₄₅ Proliferating clear cell adenofibroma
☐ ₄₆ Proliferating clear cell cystadenofibroma
☐ ₄₉ Benign Brenner tumor, typical
☐ ₅₀ Metaplastic Brenner tumor
☐ ₅₁ Proliferating Brenner tumor

☐ ₈₈ Other:

Tumor Grade:

☐ _a well differentiated ☐ _b moderately differentiated ☐ _c poorly differentiated

FIGO Stage:

<input type="checkbox"/> ₀₁ IA	<input type="checkbox"/> ₀₄ IIA	<input type="checkbox"/> ₀₇ IIIA	<input type="checkbox"/> ₁₀ IVA
<input type="checkbox"/> ₀₂ IB	<input type="checkbox"/> ₀₅ IIB	<input type="checkbox"/> ₀₈ IIIB	<input type="checkbox"/> ₁₁ IVB
<input type="checkbox"/> ₀₃ IC	<input type="checkbox"/> ₀₆ IIC	<input type="checkbox"/> ₀₉ IIIC	<input type="checkbox"/> ₁₂ IVC

POCRC - Specimen Histology Report II

Patient ID: _____ Date of analysis: ____/____/____
Form completed by: _____

A	Site: _____	B	Site: _____	Site
	Path. dx.: _____		Path. dx.: _____	1 Primary ovarian tumor
	Necrosis: _____%		Necrosis: _____%	2 Contralateral ovary - NL
	Normal cells: _____%		Normal cells: _____%	3 Metastatic tumor
Infiltr. by inflammatory cells: _____%		Infiltr. by inflammatory cells: _____%		4 Non-ovarian tissue - NL
				5 Not known
				6 Ovarian tissue - NL
				7 Tube - NL
				8 Uterus
				9 Other (specify)
C	Site: _____	D	Site: _____	
	Path. dx.: _____		Path. dx.: _____	
	Necrosis: _____%		Necrosis: _____%	Differentiation
	Normal cells: _____%		Normal cells: _____%	a well differentiated
Infiltr. by inflammatory cells: _____%		Infiltr. by inflammatory cells: _____%		b moderately differentiated
				c poorly differentiated

Pathology Diagnosis

- | | | |
|-------------------------------|----------------------------------|---------------------------|
| Non-neoplastic lesions | 1 Inadequate | 6 Benign cyst/paraovarian |
| | 2 Necrosis only | 7 Functional cyst |
| | 3 Normal ovarian or tubal tissue | 8 Corpus luteum |
| | 4 Normal fibrovascular tissue | 9 Inflammatory lesion |
| | 5 Normal other (specify) | 10 Endometriosis |
- Epithelial Tumors**
- | | |
|--|--|
| Serous tumors, benign | Serous tumors, malignant |
| 11 Serous cystadenoma | 16 Serous carcinoma of LMP |
| 12 Serous adenofibroma | 17 Serous carcinoma |
| 13 Serous cystadenofibroma | 19 Serous cystadenofibroma |
| 14 Proliferating serous adenofibroma | 20 Serous adenofibroma |
| 15 Proliferating serous cystadenofibroma | |
| Mucinous tumors, benign | Mucinous tumors, malignant |
| 25 Mucinous cystadenoma | 28 Mucinous adenocarcinoma of LMP |
| 26 Mucinous adenofibroma | 29 Mucinous carcinoma |
| 27 Mucinous cystadenofibroma | 30 Malignant mucinous adenofibroma |
| | 31 Malignant mucinous cystadenofibroma |
| Endometrioid tumors, benign | Endometrioid tumors, malignant |
| 32 Endometrioid cystadenoma | 37 Endometrioid carcinoma of LMP |
| 33 Endometrioid adenofibroma | 38 Endometrioid adenocarcinoma |
| 34 Endometrioid cystadenofibroma | |
| 35 Proliferating endometrioid adenofibroma | |
| 36 Proliferating endometrioid cystadenofibroma | |
| Mesodermal mixed tumors | 41 Mesodermal (mullerian) mixed tumor, homo. |
| 39 Benign adenofibroma | 42 Mesodermal (mullerian) mixed tumor, hetero. |
| 40 Malignant adenosarcoma | |
| Clear cell tumors, benign | Clear cell tumors, malignant |
| 43 Clear cell adenofibroma | 47 Clear cell carcinoma of LMP |
| 44 Clear cell cystadenofibroma | 48 Clear cell carcinoma |
| 45 Proliferating clear cell adenofibroma | |
| 46 Proliferating clear cell cystadenofibroma | |
| Brenner tumors, benign | Brenner tumors, malignant |
| 49 Benign Brenner tumor, typical | 52 Brenner tumor of LMP |
| 50 Metaplastic Brenner tumor | 53 Malignant Brenner tumor |
| 51 Proliferating Brenner tumor | |
| Other | 56 Unclassified epithelial tumor |
| 54 Undifferentiated carcinoma | 57 Neoplastic other (specify) |
| 55 Adenocarcinoma, NOS | |
| 98 Non-neoplastic other (specify) | 99 Other (specify) |

ORCHID – Clinical Data Form

This form should be completed 1 to 2 weeks following a participant's surgery; this allows time for all surgical and pathology reports to be submitted to her medical records file.

UPN: _____

Form completed by: _____

Name: _____

Physician ID: _____

Med. records ID: _____

Location of records: ☐₁ PGS ☐₂ UW Gyn. Onc.

I. Presenting symptoms & duration

☐ H&P not in clinic records

Symptom	Report of symptom during history	Symptom duration (weeks)			Not noted in H & P
		<4	4-8	>8	
1. Pain	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
2. Distention	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
3. Bleeding	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
4. Fatigue	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
5. Dyspepsia	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
6. Weight change	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
7. Bladder changes	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
8. Bowel changes	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄
9. Other: _____	<input type="checkbox"/> ₁ Neg. <input type="checkbox"/> ₂ Pos. → Duration?	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄

Comments: _____

II. Pre-operative CA 125 screens

☐ None listed in clinic records

Date of exam	Results		
_____	_____ U/ml	<input type="checkbox"/> ₁ Dynacare	<input type="checkbox"/> ₂ Other laboratory
_____	_____ U/ml	<input type="checkbox"/> ₁ Dynacare	<input type="checkbox"/> ₂ Other laboratory
_____	_____ U/ml	<input type="checkbox"/> ₁ Dynacare	<input type="checkbox"/> ₂ Other laboratory
_____	_____ U/ml	<input type="checkbox"/> ₁ Dynacare	<input type="checkbox"/> ₂ Other laboratory

III. Size of ovarian mass Abstract from pathology reports.

Date of report	LT	RT	Bidimensional tumor size (note units)	
_____	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	_____ x _____	_____
_____	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	_____ x _____	_____

IV. Post-operative diagnosis

Record all that apply.

	Ca	LMP	Ben	Nml	Dx.:
Rt. ovary	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	_____
Lt. ovary	<input type="checkbox"/> ₁	<input type="checkbox"/> ₂	<input type="checkbox"/> ₃	<input type="checkbox"/> ₄	_____

Other relevant conditions: _____ Dx.: _____

_____ Dx.: _____

_____ Dx.: _____

Appendix K
Histology Reports

Overall										Stage of disease in patient consented tissue																						
										Stage 1		Stage 2		Stage 3		Stage 4		Unknown														
										Women Represented	Blood Vials Collected	Blood Vials Available	Tissues Collected	Tissues Available	Primary Tissues	Women Represented	Tissues Collected	Tissues Available	Women Represented	Tissues Collected	Tissues Available											
Description of Specimens Collected (excludes fluids)																																
All Studies - Specimens from Patients without cancer DX																																
All Specimens										235	2057	1850	1558	1483	314	307	1	5	5	1	14	14	233	1539	1464							
											2057	1850	1558	1483	314	307		5	5		14	14		1539	1464							
ORCHID - Diseased Specimens from OVCA Patients																																
Serous										53	504	291	484	402	254	204	2	15	13	1	12	11	43	394	327	7	63	51				
Mucinous										3	23	13	15	8	15	8	3	15	8													
Endometrioid										4	57	43	22	18	22	18	4	22	18													
Mesodermal mixed										2	12	9	20	18	15	14				1	10	9										
Clear cell tumors										5	54	39	30	28	24	22	2	12	10				1	10	9							
Other -- Undifferentiated carcinoma										1	9	4	2		2																	
Other -- Adenocarcinoma, NOS										18	139	93	128	112	65	57																
Other -- Unclassified epithelial tumor										4	42	25	31	20	17	10																
Other -- Neoplastic										1	14	12	5	5	5	5				1	6	5	1	7	6	2	18	9	1	2	1	
											854	519	737	611	419	338		11	64	49	3	28	25	66	551	465	11	92	71	1	2	1
ORCHID - Disease specimens from LMP Patients																																
Serous										8	77	47	37	28	35	26		6	29	22	2	8	6									
Mucinous										2	29	19	11	11	11	11		2	11	11												
											106	66	48	39	46	37			40	33		8	6									
ORCHID - Benign Specimens from OVCA Patients																																
Non-neoplastic lesions										3	34	27	8	8				2	6	6	1	2	2									
Serous										2	28	19	5	5	5			1	2	2				1	3	3						
Mucinous										1	3	2	2	2	2	2		1	2	2												
Other -- Benign										1	7	4	1	1	1			1	1	1												
											72	52	16	16	2	2		5	11	11	1	2	2	1	3	3						
ORCHID - Normal Specimens from OVCA Patients																																
Non-neoplastic lesions										59	573	352	166	145	13	11		14	46	40	4	10	8	33	90	78	7	15	14	1	5	5
											573	352	166	145	13	11			46	40		10	8		90	78		15	14		5	
ORCHID - OVCA patient specimens of unknown pathology																																
Non-neoplastic lesions										1	19	15	2	2																		
											19	15	2	2																		

[illegible]

Appendix L
Pull List

Pull List Ordered by both Source Box and Target Box

Ordered by source box number

id	specimenid	box	ncol	newboxnm	newcol
117	100775	KIVI - SER 1 9 8	Nora1	3 3	
731	205444	KIVI - SER 5 8 9	Nora1	4 7	
164	100593	KIVI - SER 6 8 7	Nora1	3 2	
745	205463	KIVI - SER 7 7 4	Nora1	4 5	
152	100613	KIVI - SER 7 1 9	Nora1	4 6	
201	200089	KIVI - SER 8 7 8	Nora1	3 1	
229	201139	KIVI - SER 9 8 5	Nora1	2 9	
228	200988	KIVI - SER 10 2 8	Nora1	2 8	
227	201045	KIVI - SER 11 9 3	Nora1	4 4	
279	200749	KIVI - SER 12 4 8	Nora1	4 3	
286	200789	KIVI - SER 13 2 3	Nora1	4 2	
732	205415	KIVI - SER 14 4 6	Nora1	2 6	
292	201375	KIVI - SER 14 4 5	Nora1	2 7	
748	205560	KIVI - SER 15 6 7	Nora1	2 5	
327	202500	KIVI - SER 16 3 7	Nora1	2 4	
346	200219	KIVI - SER 18 1 1	Nora1	4 1	
388	202238	KIVI - SER 19 7 4	Nora1	2 1	
366	202594	KIVI - SER 19 4 6	Nora1	2 2	
362	202046	KIVI - SER 19 1 1	Nora1	2 3	
357	200911	KIVI - SER 20 6 4	Nora1	1 9	
386	202213	KIVI - SER 23 3 2	Nora1	1 7	
412	202989	KIVI - SER 23 2 3	Nora1	1 8	
429	203323	KIVI - SER 25 7 6	Nora1	3 9	
477	203383	KIVI - SER 26 4 6	Nora1	1 6	
576	204357	KIVI - SER 33 1 9	Nora1	3 8	
595	204391	KIVI - SER 34 7 3	Nora1	1 4	
563	204257	KIVI - SER 34 5 3	Nora1	1 5	
640	204503	KIVI - SER 37 3 3	Nora1	1 3	
632	204679	KIVI - SER 38 4 5	Nora1	3 7	
709	205232	KIVI - SER 41 4 5	Nora1	3 5	
696	205169	KIVI - SER 41 3 2	Nora1	3 6	
755	205691	KIVI - SER 42 9 5	Nora1	3 4	
760	205784	KIVI - SER 43 5 7	Nora1	1 1	
697	205758	KIVI - SER 43 2 6	Nora1	1 2	

Ordered by target box number

id	specimenid	box	ncol	newboxnm	newcol
760	205784	KIVI - SER 43 5 7	Nora1	1 1	
697	205758	KIVI - SER 43 2 6	Nora1	1 2	
640	204503	KIVI - SER 37 3 3	Nora1	1 3	
595	204391	KIVI - SER 34 7 3	Nora1	1 4	
563	204257	KIVI - SER 34 5 3	Nora1	1 5	
477	203383	KIVI - SER 26 4 6	Nora1	1 6	
386	202213	KIVI - SER 23 3 2	Nora1	1 7	
412	202989	KIVI - SER 23 2 3	Nora1	1 8	
357	200911	KIVI - SER 20 6 4	Nora1	1 9	
388	202238	KIVI - SER 19 7 4	Nora1	2 1	
366	202594	KIVI - SER 19 4 6	Nora1	2 2	
362	202046	KIVI - SER 19 1 1	Nora1	2 3	
327	202500	KIVI - SER 16 3 7	Nora1	2 4	
748	205560	KIVI - SER 15 6 7	Nora1	2 5	
732	205415	KIVI - SER 14 4 6	Nora1	2 6	
292	201375	KIVI - SER 14 4 5	Nora1	2 7	
228	200988	KIVI - SER 10 2 8	Nora1	2 8	
229	201139	KIVI - SER 9 8 5	Nora1	2 9	
201	200089	KIVI - SER 8 7 8	Nora1	3 1	
164	100593	KIVI - SER 6 8 7	Nora1	3 2	
117	100775	KIVI - SER 1 9 8	Nora1	3 3	
755	205691	KIVI - SER 42 9 5	Nora1	3 4	
709	205232	KIVI - SER 41 4 5	Nora1	3 5	
696	205169	KIVI - SER 41 3 2	Nora1	3 6	
632	204679	KIVI - SER 38 4 5	Nora1	3 7	
576	204357	KIVI - SER 33 1 9	Nora1	3 8	
429	203323	KIVI - SER 25 7 6	Nora1	3 9	
346	200219	KIVI - SER 18 1 1	Nora1	4 1	
286	200789	KIVI - SER 13 2 3	Nora1	4 2	
279	200749	KIVI - SER 12 4 8	Nora1	4 3	
227	201045	KIVI - SER 11 9 3	Nora1	4 4	
745	205463	KIVI - SER 7 7 4	Nora1	4 5	
152	100613	KIVI - SER 7 1 9	Nora1	4 6	
731	205444	KIVI - SER 5 8 9	Nora1	4 7	

***“Use of Novel Technologies to Identify and Investigate Molecular
Markers for Ovarian Cancer Screening and Prevention”***

**TISSUE COLLECTION, PROCESSING AND TRANSPORT PROTOCOL
for
ORCHID CORE**

Ovarian Research Collaboration Helping to Improve Detection

Conducted by Investigators at Fred Hutchinson Cancer Research Center (FHCRC), University of Washington (UW) and Virginia Mason Research Center (VM). Funding by the Marsha Rivkin Center for Ovarian Cancer Research and the United States Army Medical Research and Materiel Command (USAMRMC).

Dates of Study: February 1998 – September 2000

Principal Investigator:

Nicole D. Urban, ScD
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North - MP-804
PO Box 19024
Seattle, WA 98109
(206) 667-4677

Investigators:

Garnet Anderson, PhD (FHCRC)
Nancy Kiviat, MD (UW)
Charles Drescher, MD (FHCRC)
Mary Anne Rossing, PhD (FHCRC)
Jane Kuypers, PhD (UW)
Leona Holmberg, MD (FHCRC)

Location(s) of Study:

Fred Hutchinson Cancer Research Center (FHCRC)
1100 Fairview Avenue North
Seattle, WA 98109
(206) 667-5000

Pacific Gynecology Specialists (PGS)
1101 Madison Street, #1500
Seattle, WA 98104
(206) 215-6595

Marsha Rivkin Center for Ovarian Cancer Research
1221 Madison Street, Suite 1410
Seattle, WA 98104
(206) 386-2419

University of Washington (UW)
School of Medicine
Seattle, WA 98195
(206) 543-2100

Table of Contents

- 1.0 Selection Criteria
 - 1.1 Recruitment
 - 1.2 Data Collection, Flow and Subject Identification
 - 1.3 Risks to Subject
 - 1.4 Follow-up Procedures
 - 1.5 Medical Monitor
- 2.0 Specimen Collection
 - 2.1 Tissue Collection
 - 2.2 Blood Collection
 - 2.3 Specimen Storage
 - 2.4 GOG/External Specimens
 - 2.5 Specimen Allocation
- 3.0 Specimen Processing Facility (Laboratory Core)
 - 3.1 Equipment
 - 3.2 Laboratory Supplies
 - 3.3 Labeling
 - 3.4 Computerized Specimen Tracking System
 - 3.5 Specimen Transport Preparation
- 4.0 Specimen Characterization and Analysis
 - 4.1 Immunohistochemistry
 - 4.2 PCR-SSCP for p53 Mutations
 - 4.3 Preparation of RNA from Peripheral Blood Samples for RT-PCR Assays
 - 4.4 Radioimmunoassay for CA-125
- 5.0 Safety Procedures
- 6.0 Disposition of Data
- 6.1 Biostatistical Reviews
- 7.0 Protocol Modification
 - 7.1 Reporting of Serious and Unexpected Adverse Events
- 8.0 Use of Information Arising from Study
- 9.0 Personnel to Conduct Project
- 10.0 Signature of Principal Investigator and Institution Approval

Introduction

Blood and tissue specimens for this research will be obtained from consenting patients identified by physicians practicing at Pacific Gynecology Specialists (PGS) and at the University of Washington (UW), and performing ovarian related surgery on the campuses of Swedish Medical Center and the University of Washington Medical Center.

Tissue collection technicians will be available during patient surgery to collect the appropriate tissue and serum samples for initial transport to the pathology laboratory, and later to the study laboratory/repository for further analysis and storage. Specimens will be placed in freezers in a pre-defined location, pending processing into the inventory control system. The specimen tracking subsystem will uniquely identify each separate item entered into the repository, document its location in the freezer, and its disposition. The transfer of specimens between sites of study will be tracked using a computerized data tracking system. Specimens that are not transported to project laboratories for analyses will remain in storage at the Core repository in a -70°C freezer or a liquid nitrogen freezer.

After characterization of specimens, requests for access to specimens by both Project Investigators and non-Project Investigators will be reviewed. These requests will be granted in the form of defined number of specimens of each type, from a selection criteria defined by final diagnosis and potentially other design criteria. Upon request approval, the specimens will be retrieved and delivered on dry ice to the Project Investigator. Please see section 2.5 about specimen allocation to non-Project Investigators.

1.0 Selection Criteria

Tissue and blood specimens for this research project will be obtained from women undergoing surgery for ovarian-related disorders at Swedish Hospital Medical Center by Pacific Gynecology Specialists surgeons or from women identified through the gynecological oncology service at the University of Washington Medical Center.

Over a period of two years, 500 women with appropriate diagnoses (70 ovarian cancer cases, 430 women with benign disease, or with no ovarian abnormalities) will be recruited for this research study. The age range of the participants will be between 20-80 years old. Only women undergoing ovarian related surgery will be invited to participate in the tissue and serum collection component of the research study. Minority representation will be reflective of the representation seen at PGS and the UW.

1.1 Recruitment

All appropriately identified PGS and UW patients scheduled to undergo surgery for ovarian related disorders, will be considered as potential candidates for this research study. These patients will be invited to participate in the study by the attending physician or nurse at the time of the pre-operative office visit. Interested patients will be provided written information describing the research study to review. Should the patient choose to participate in this protocol, the attending physician, nurse or a research study staff member will review the consent form and other required enrollment documents with the patient.

Patients will be given an opportunity to ask questions and address any concerns they may have about participating. Patients will be informed that the decision to participate will not affect their treatment in any way, and that in agreeing to participate, they reserve a right to terminate their participation at any time without prior notice.

1.2 Data Collection, Flow and Subject Identification

Women who consent to participate will be given a study enrollment packet. This packet includes a study letter or brochure, a participant enrollment form, a medical records release form, and a consent form to be completed at the pre-operative office visit. A self-administered 20-minute questionnaire with a self-addressed, stamped envelope will also be included in the enrollment packet. The patient may complete the questionnaire during the pre-operative office visit, or return it by mail at a later date. All materials will be pre-labeled with a unique packet identification number.

Clinic or research study staff will complete the portion of the patient enrollment form reserved for internal use only. This portion of the patient enrollment form indicates the date of scheduled surgery, institution of enrollment and identification of the enrolling physician, other medical or research study personnel. The enrolling staff member will ensure that the patient has fully completed the informed consent, and will note in the patient's chart that she has been approached and has agreed to participate in the research study. To ensure that this patient will not be approached again about the study, the enrolling staff will be responsible for noting patient participation or refusal in the patient's medical chart.

The completed enrollment forms and two copies of the signed informed consent will be sent to the Clinical, Statistical and Laboratory Coordination Core. The Data Coordinator will enter data from the enrollment forms into the study database on a daily basis. At the time that data from enrollment forms is entered into the study database, each participant will be given a unique participant number (UPN). All self-administered questionnaires, after completion by the participants, will be returned to the Core facility for editing and data entry.

The UPN will be used to label all data collection forms, requisition forms, and transport forms. A unique 6-digit number will be used to label all specimens, with a duplicate of

the label attached to the specimen collection form. The unique label number for each specimen will be linked to each participant's UPN in the specimen inventory database.

1.3 Risks to Subject

All tissue that is obtained for the purposes of this research study is collected only after it has been removed for the purposes of the surgical procedure that the participant is undergoing. All blood (no more than 40 cc) obtained for the purposes of this research study is collected by the anesthesiologist prior to or during the actual surgical procedure. This collection will not pose any additional risks to the participant. The participant may experience potential discomfort by not being given the results of the analyses to be conducted by the research project.

Any precautions possible will be taken to ensure that the participant's risks are minimized or eliminated. These include extensive precautions to maintain the confidentiality of all study records identifying patient information by enforcing and following strict protocols. These procedures include a pledge of confidentiality by all study personnel at Fred Hutchinson Cancer Research Center, Pacific Gynecology Specialists and the University of Washington, data handling procedures, network and password protection, and proper storage and handling of all files and specimens.

Study participants are informed that their personal identity will not be revealed in any publication or release of results. Study participants are also informed that representatives from the U.S. Army Medical Research and Materiel Command will have access to their study records, and may inspect the records of the research in their duty to protect human subjects in research.

1.4 Follow-up Procedures

1.4.a. Incomplete Enrollment

With adherence to stringent enrollment procedures, very little participant follow-up for this study is anticipated. However, follow-up may be required if a patient has not fully completed enrollment forms or has not returned the study questionnaire. In such situations, a written request, followed by one telephone call, will be made by the Study Coordinator. A letter will be mailed to the participant if their enrollment materials are incomplete or if their questionnaire has not been received by the Core within thirty days of their enrollment. After fourteen days, a follow up call will be made to the participant if she has not responded to the request. The call script included in the appendix of this protocol will be employed to inquire about the questionnaire or clarify ambiguous or incomplete information on the enrollment forms.

1.5 Medical Monitor

A medical monitor has been assigned to this research study. The medical monitor is Dr. Saul Rivkin of the Swedish Hospital Medical Center Tumor Institute. Dr. Rivkin is an exceptionally qualified physician, who is not associated with the protocol. Dr. Rivkin is fully able to provide medical care to the research subjects for conditions that may arise during the conduct of the study. A short biosketch and Dr. Rivkin's Curriculum Vitae is included with this protocol.

2.0 Specimen Collection

Each day, a report will be generated showing the most up-to-date information on scheduled surgeries for study participants. This report includes the patient name and UPN, surgeon, date, time and location of surgery.

The schedule and amount of specimen to be collected will vary throughout the study. It is anticipated that three to five collections from qualified participants will be conducted per week. The amount of tissue specimen collected for the purposes of the research study will also vary from 1 gram up to 5 grams. Only that which is not needed for the purposes of pathologic diagnosis will be available for ORCHID study collection. This will be determined by the clinical pathologist on a case by case basis.

For each scheduled surgery, a packet containing the UPN specimen collection and processing forms, and a copy of the informed consent, will be assembled and sent to the specimen collection team with the scheduled surgery report. In addition, the collection team will be provided with a pre-assembled specimen kit for tissue and blood collection. The Tissue Collection Specialist will be responsible for ordering, maintaining, and assembling supplies for the specimen kit. The specimen kit will include the following pre-labeled items:

- Three (3) biohazard bags (with foil) for snap frozen primary and metastatic tumor and normal specimens
- One (1) truncated embedding mold for primary tumor/tissue specimens frozen in OCT compound
- Three (3) 15 ml. formalin jars for fixed specimens
- One (1) STM tube for primary tumor tissue
- Two (2) 5 ml. lavender-top EDTA tubes for blood collection
- Three (3) 10 ml. red-top tubes for blood collection

- Ten pre-labeled cryovials for serum, plasma, and WBC pellet collection
- 4 lbs. dry ice
- Biohazard stickers and dry ice labels

2.1 Tissue Collection

The Tissue Collection Specialist, who maintains a log of scheduled surgeries, will work with operating room physicians and personnel to notify them prior to the beginning of a participating patient's surgery. During the entire surgical procedure, the attending surgeon and surgical personnel will be responsible for monitoring the patient's vital signs and condition.

Immediately after the surgeon has removed the necessary tissue and the pathologist has taken what is required for pathologic diagnosis, the Tissue Collection Specialist will be allowed to collect specimens from this removed tissue for the purposes of this study. The tissue samples will be processed according to the guidelines below:

Ovarian Tissue:

- Surgical specimens will be placed in labeled sterile containers containing 0.9% sodium chloride and transported by Tissue Collection Specialist into the processing area located in the frozen section room.
- Under the direction of a clinical pathologist, tissue necessary for clinical evaluation will be removed.
- Tissue used for the proposed studies will be selected from an area representative of the specimen and as free of necrosis as possible.
- In the case of normal ovaries, the surface epithelium will be manually scraped from the ovary and snap frozen, to minimize contaminating stromal tissue.

Frozen Tissue Amounts and Preparation:

- A minimum of 1gm and up to 5 gm of tissue, which will be divided into approximately 1 cm³ sections. Each section will be completely wrapped in aluminum foil and immersed in liquid nitrogen for a minimum of 3 minutes.
- Frozen tissues will then be placed in biohazard bags pre-labeled with the UPN and tissue type.
- Specimens will be stored on dry ice for transport to the core facility.
- For the OCT mold, truncated molds will be pre-labeled with the UPN using a SECURLINE permanent marker.

- Each mold will be partially filled with OCT medium and pre-cooled by holding over (not in) liquid nitrogen until OCT medium loses transparency.
- Approximately 1 gm of tissue will be placed in the mold, covered with OCT medium and immersed into liquid nitrogen until completely solid.
- Specimens will be placed into a UPN labeled biohazard bags and stored on dry ice for transport to the core facility.

Paraformaldehyde-Fixed Tissue Amounts and Preparation:

- A portion of tumor smaller than or equal to 1x1 cm and no thicker than 2 mm will be selected and placed in cold 4% paraformaldehyde and stored for 2 hours at 4°C.
- After a 2 hour fixation, the 4% paraformaldehyde will be discarded and replaced with cold 30% sucrose, and the sample will be stored at 4°C.
- Tissue will initially float in sucrose but when left overnight will sink.
- After the tissue has sunk, but no longer than 24 hours after fixation, the tissue will be imbedded in OCT as described in protocol for tissue preparation. (see above)
- The mold will be placed into a labeled biohazard bag and stored at -70°C until transfer to the liquid nitrogen freezer at the Core laboratory/repository .

2.2 Blood Collection

Prior to each surgery, patient consent will be obtained to collect blood. The research collection team will work with the anesthesiologist and notify him/her prior to the beginning of surgery that a patient is participating in the research study.

Blood Collection and Preparation:

- All requisite serum collection vials and clot tubes will be prepared and labeled with the patient name, UPN, and date of collection prior to surgery.
- At the time of surgery, the anesthesiologist will collect up to 30 cc of whole blood in a non-heparinized, red-top tubes using a vacutainer and 21 ga needle just prior to the surgeon removing tissue samples.
- Blood will be allowed to stand for 30-120 minutes and then stored in the operating room on ice until it is transferred to the Core laboratory/repository.
- At the Core laboratory, the blood will be placed into a refrigerated centrifuge and spun for 10 minutes at 2500 rpm at 4°C.
- After centrifugation, the vials will be placed in styrofoam container holding ice and placed under a hood. The tops of the vials will be swabbed with alcohol and the serum will be removed using a sterile 21 gauge needle and syringe. The serum is then aliquoted into study-labeled 250 ul NUNC tubes.

- NUNC tubes will then be placed into a UPN labeled plastic baggie and stored on ice for transport to the Core laboratory/repository.

WBC Pellet Preparation:

- The anesthesiologist will collect an additional 10 cc of whole blood in lavender-top EDTA tubes using a vacutainer and 21 ga needle just prior to the surgeon removing tissue samples.
- Blood will be stored in the operating room on ice until it is transferred to the Core laboratory within six hours of collection.
- At the laboratory, 0.5 ml of this blood will be added to a 2.0 ml Sarstedt tube. Next, 1 ml of specimen wash solution is added to this tube to lyse the RBCs.
- The Sarstedt tube will then be centrifuged to pellet the WBCs.
- After the WBC pellets have been washed three times with wash solution, the pellets will be transferred to a prelabeled cryovial.
- These WBC pellet cryovials will then be placed in the -70° C for long-term storage.

2.3 Specimen Storage

All tissue and sera obtained by the specimen collection team will be placed on dry ice for transport to the Core facility. The Tissue Collection Specialist will transport all samples to the Core laboratory the same day as collected. Upon receipt at the Core, all frozen tissue specimens will be stored in a liquid nitrogen freezer. Serum, plasma and WBC pellets will be stored in a -70° C freezer.

2.4 GOG/External Specimens

Additional tissue specimens provided by the GOG will be treated as collected specimens. Upon receipt of frozen specimens (shipped overnight on dry-ice by the GOG), each patient will be assigned a PIN (unique PIN will be allocated to GOG samples) and all specimens will be stored in the liquid nitrogen freezer prior to processing and characterization. Any information accompanying the specimens such as date of collection, age of patient at time of surgery, pathology and histology information, and other non-identified demographic data will be entered into the tracking system.

2.5 Specimen Allocation

After characterization in the Laboratory Core, specimens will be made available to Project Investigators. After Project needs have been met, specimens may be made available to non-Project Investigators. In such circumstances, the non-Project

Investigators will be required to complete a review process for use of said specimens. All specimens transferred to non-Project Investigators must receive approval and/or certification from Study Investigators, and the FHCRC IRO. Specimens provided to commercial entities, or Investigators in collaboration with a commercial entities must also receive approval from the FHCRC Human Specimens Committee.

Non-Project Investigators and/or commercial entities will be asked to submit a proposal to this study's Investigators, stating the following: 1) the hypothesis to be tested 2) how the specimens will be used, 3) the amounts and types of specimens requested and 4) preliminary data. In addition, a biostatistical consult will be conducted to ensure that sample sizes are sufficient and that the study is sound in design.

If approved by study Investigators, non-Project Investigator(s) and/or commercial entities will be required to submit an Institutional Review Board application to the FHCRC IRO for research protocol review and approval.

If approved by the FHCRC IRO, and if not a commercial entity, or an investigator(s) involved in a collaboration with a commercial entity, the specimen request will be considered approved. If approved by the FHCRC IRO, and if a commercial entity, or an investigator(s) in a collaboration with a commercial entity, non -Project Investigators will be required to submit application to the FHCRC Human Specimens Committee for research protocol review and approval.

In either situation, upon request approval, the specimens will be retrieved and delivered on dry ice to the non-Project Investigator(s). The Project Coordinator, Suepattra May, will serve as the Repository Gatekeeper and will ensure that specimens and/or corresponding data are provided only to Investigators that are in full compliance with the application protocol. In addition, Ms. May will be responsible for all application materials and other paperwork associated with this process, including completed Confidentiality Pledges.

3.0 Specimen Processing Facility (Laboratory Core)

All tissue specimens will undergo processing at the Core laboratory facility before long-term storage and/or transport to project or non-project laboratories. All red top blood specimens will undergo processing at the Dynacare Laboratory of Pathology Stat Laboratory on the Swedish Medical Center Campus or at the Core laboratory facility before long-term storage and/or transport to project or non-project laboratories. All purple top (EDTA) blood products will undergo processing at the Core laboratory facility before long-term storage and/or transport to project or non-project laboratories

3.1 Equipment

The laboratory facility will be equipped with all necessary specimen processing equipment and supplies. Equipment for this facility includes a refrigerator, two freezers, MicroProbe IHC system, microcentrifuge, PCR hood, thermocycler, hybridization oven, gel electrophoresis tank and gel dryer, vacuum pump and trap.

The Research Technician will conduct all quality assurance of laboratory equipment and arrange for routine maintenance of equipment as recommended by manufacturers. All monitoring and maintenance checks of equipment will be recorded in individual maintenance logs. Freezer temperatures will be monitored daily using a thermometer linked to an alarm system and recorded daily by the Research Technician.

Each of the freezers will be equipped with an eight-hour CO₂ back-up system and a temperature-sensitive alarm system that alerts the building maintenance staff when the interior temperature reached a designated temperature. The CO₂ tank will be checked on a monthly basis to ensure that it has not been emptied. Each month, the alarm system will be tested to ensure that it will sound should the freezer temperature rise above -50°C. The Project Manager will be available for monitoring the freezers during non-office hours. In the case of equipment failure, a back-up freezer space will be available for specimens.

3.2 Laboratory Supplies

The Core will maintain, at minimum, a two-month inventory of specimen collection, processing and transport supplies. The Research Technician and Tissue Collection Specialist will be responsible for ensuring that the minimum requisite levels of supplies are available for the researchers and collection team.

3.3 Labeling

A number of labels specifying a unique specimen identification number will be used for collection and processing of each participant's tissue and blood specimen for the duration of the study. Labels will be used for all collection and processing forms, specimen containers and vials, and logs.

The freezer boxes for each specimen type will be labeled accordingly. Freezer boxes will be pre-labeled with appropriate specimen characterizations, so that as new specimen types are received, the specimens will be added to the appropriate freezer box.

3.4 Computerized Specimen Tracking Program

A computer database will serve as the specimen inventory system that tracks specimens collected by the Tissue Collection Specialist. This system will uniquely identify each separate item entered into the repository, document its location in the freezer, and its disposition as it is transferred to project laboratories for analysis.

3.5 Specimen Transport Preparation

After the specimen requisition process has been completed, the Research Technician will prepare the specimens for transport to the project laboratories. To ensure the integrity of the specimens, the freezer boxes will not be removed from the freezer for processing until all transport supplies are available for performing the transport procedure. The specimens will be packaged in a styrofoam box according to the following procedure:

- A 2" layer of pelleted dry ice nuggets will be placed on the bottom of the styrofoam box;
- The freezer boxes will be placed in a self-sealed, or waterproof sealed plastic bag;
- A biohazard symbol will be affixed to the outside of the plastic bag;
- The sealed plastic bags will be placed in the styrofoam box on top of the dry ice;
- Another 2 lbs. of pelleted dry ice will be layered on top of and around the plastic bags;
- Any empty spaces will be stuffed tightly with newspaper;
- The seams of the styrofoam box will be taped with waterproof tape. No scotch or masking type will be used;
- The styrofoam container will then be placed inside the transport carton;
- The transport carton will be labeled "Diagnostic Specimen" with a grease pen;
- The transport carton will be temporarily stored in the freezer until it is picked up for transport.

If it is noted that one of the vials or containers is cracked, the cracked container will be placed into another larger vial or container and may be transported separately in a sealable plastic bag, at the discretion and evaluation of the laboratory director.

The Research Technician or Data Coordinator will advise the appropriate project laboratory that a shipment of specimen is on its way. The project laboratory personnel will contact the Core Repository staff to inform them that the shipment has been received.

4.0 Specimen Characterization and Analysis

All tissues will be reviewed by the Core facility prior to their analyses in Projects 1 and 2. The review will include the frozen section of the biopsy material that has been submitted for further studies in Projects 1 and 2. For "normal" tissue, multiple levels of tissue submitted as normal will be reviewed. Tumors will be classified according to the WHO classification system as:

- 1) an epithelial tumors including serous (benign, borderline, malignant), mucinous (benign, borderline, malignant), endometrioid, clear cell, Brenner, mixed epithelial tumors, undifferentiated carcinomas, or unclassified epithelial tumors);
- 2) a sex cord-stromal tumors;
- 3) a lipid cell tumors;
- 4) a germ cell tumors;
- 5) a gonadoblastomas;
- 6) a soft tissue tumors (not specific to ovary);
- 7) an unclassified tumors, a metastatic tumors or
- 8) a tumor-like conditions.

All formalin-fixed, paraffin-embedded biopsy tissues will be examined by immunohistochemical techniques for the presence of the oncoproteins HER2/neu, Myc and for the intranuclear accumulation of mutant p53 proteins. DNA will also be extracted from fresh tissue and will be tested by PCR-SSCP analysis to screen for mutations in the p53 DNA. These IHC and mutation analyses will be performed utilizing the following protocols.

4.1 Immunohistochemistry (IHC)

- For all IHC assays, the Elite Vectastain ABC kit (Vector Laboratories) will be used with a specific primary antibody for each protein to be assayed.
- Sections four to six microns thick will be cut, mounted on silanated slides, dewaxed, and rehydrated.
- After blocking, the primary antibody will be added.
- The slides will then be washed and the secondary antibody, a biotinylated goat anti-mouse IgG, will be added.
- After a second wash, peroxidase-conjugated avidin will be added, washed and then reacted with the chromogen diaminobenzidine.
- Finally, the slides will be counter-stained with methyl green. Normal tissue will be stained as a negative control. An antibody against keratin AE1/AE3 (Boehringer Mannheim) will be used as a positive control.
- The intensity of staining of the cases will be determined (0 = not greater than the negative control, 1+ = light staining, 2+ = moderate staining, 3+ = heavy staining) and compared to the intensity of staining of the normal ovarian tissue.

4.1 PCR-SSCP for p53 Mutations

- All cases will be screened for mutations in the p53 gene by single-strand conformation polymorphism (PCR-SSCP) analysis.

- DNA will be purified from fresh tissue by Proteinase K digestion, phenol/chloroform extraction and ethanol precipitation.
- Mutations in exons 5-9 will be detected by means of PCR-SSCP analysis (4), amplifying 0.1-1.0 µg of DNA in separate reaction mixes with primer pairs for exons 5, 6, 7, and 8.
- The amplification products will be denatured and run on a polyacrylamide gel. Bands will be visualized by a silver stain.
- To further detect mutations in exons 7-9, a fifth primer pair will be used to generate a PCR fragment containing exons 7-9 which will be digested with the restriction enzyme MspI.
- Any variant band detected by PCR-SSCP analysis which do not conform to the pattern of the common p53 mutations (which serve as positive controls for the assay) will be cut out of the gel, eluted in TE-4, and directly sequenced using dye terminator reactions (utilizing the same primer pairs) and analyzed on an ABI sequencer.

4.3 Preparation of RNA from Peripheral Blood Samples for RT-PCR Assays

- Whole blood will be collected into EDTA anticoagulant tubes and 0.5 ml aliquots will be added to 1 ml of lysis buffer containing 0.4% detergent.
- The unlysed cells will be pelleted and washed two times with the lysis buffer.
- The pellets from 2 ml of whole blood will be combined and resuspended in 1 ml of UltraSpec RNA isolation reagent (Biotecx) and the total cellular RNA will be purified according to the manufacturer's protocol.
- One to two µg of total RNA will be added to a reverse transcriptase PCR reaction using primers specific for the protein of interest and the PCR amplicons will be detected by dot blot hybridization.

4.3 Radioimmunoassay for CA-125

One 250 µl aliquot of serum from the selected women will be obtained for CA-125 detection.

- Immunoradiometric assay of CA-125 levels will be performed using the commercially available RIA kit (Centocor, Malvern, PA.).

5.0 Safety Procedures

All Core personnel handling tissue and blood specimens are required to become familiar with and adhere to the applicable sections of the Protocol for Specimen Collection, Processing and Transport. A copy of any local and state requirements relating to the collection and processing of blood products must be on file with at the Statistical, Clinical and Laboratory Coordination Core.

All blood and tissue specimens will be handled as potentially infectious material. The Core has adopted the Universal Precautions for blood collection and processing. Universal Precautions refers to an approach to infectious disease control which assumes that every direct contact with body fluids is infectious. This approach requires that persons who may be in direct contact with body fluids be protected as though all body fluids contain blood-borne pathogens. All Core clinic and laboratory personnel will be guided by the universal precautions in order to protect all persons from parenteral, mucous membrane and non-intact skin exposures to blood borne pathogens.

It must be noted that any body fluid may contain microorganisms capable of transmitting disease. Therefore, appropriate protective attire must be worn where there is potential for direct contact with any body fluid or tissue. Core personnel will be required to change gloves and wash hands after handling laboratory specimens containing body fluids.

All procedures involving blood or other potentially infectious materials must be performed in a manner which minimizes splashing, spraying and aerosolization of these substances.

Core personnel will adhere to the following guidelines as regards each topic:

- *Hand Washing* - Employees must wash their hands:
 - Immediately after contact with blood or other infectious materials (even if gloves were worn);
 - Before and after using restroom facilities;
 - After removal of gloves and/or other protective clothing;
 - Upon leaving the work area where blood or other infectious materials are present.
- *Personal Protective Equipment* - Personal protective equipment such as fluid resistant gowns, gloves, goggles, and masks must be available and used in areas where blood and or other potentially infectious materials are handled. Supplies such as face shields, head and foot coverings must be available and used when invasive procedures are being carried out.
- *Accessibility of Equipment* - Appropriate protective clothing must be worn when the employee has a potential for exposure to blood and other potentially infectious materials.

- *Removal of Equipment* - Personal protective equipment (disposable clothing) must be removed immediately upon leaving the work area and placed in a labeled infectious waste container for disposal.
- *Gloves* - The use of disposable gloves is mandatory for procedures in which body fluids or other potentially infectious materials are handled. Gloves should be changed when contaminated and prior to entering common areas (such as elevators or restrooms). Latex or vinyl gloves are appropriate. Gloves must be worn when the Core personnel has the potential for direct skin contact with:
 - Blood;
 - Infectious materials;
 - Tissue;
 - Mucous membranes;
 - When handling items or surfaces soiled with blood or other infectious materials.
- Gloves should not be used if they are peeling, cracked or discolored, or if they have punctures, tears, or other evidence of deterioration. If an employee has an open cut or abrasion on the hand(s), the area must be protected with a Band-Aid underneath the glove.
- *Gowns* - Fluid resistant gowns or aprons must be worn if there is a potential for soiling of clothes with blood or other potentially infectious materials.
- *Surgical Caps or Hoods* - Surgical caps or hoods must be worn if there is a potential for splashing or spattering of blood or other potentially infectious materials on the head.
- *Fluid Proof Shoe Covers* - Fluid-proof shoe covers must be worn if there is a potential for shoes to become contaminated with blood or other potentially infectious materials.
- *Masks, Eye Protection and Face Shields* - Masks, eye protection, or chin-length face shields must be worn whenever splash, spray, spatter, droplets or aerosols of blood or other potentially infectious materials may be generated and there is a potential for eye, nose, or mouth contamination.
- *Spill Clean-Up* - All equipment and working surfaces must be properly cleaned and disinfected after contact with blood, tissue or other potentially infectious materials. Broken glassware which may be contaminated must be removed by mechanical means, such as tongs, cotton swabs or forceps. Chemical germicides and disinfectants should be used at recommended dilutions to decontaminate all spills of blood and other potentially infectious materials. All spills must be cleaned immediately while adhering to the following guidelines:
 - Gloves must be worn when wiping up a spill;

- An appropriate disinfectant must be used; and
 - Disinfectants (at appropriate dilution) should be poured onto a paper towel for wiping up small spills.
 - No trigger type spray bottles or other equipment which would aerosolize the disinfectant should be used.
-
- *Laundry* - All laundry is assumed to be contaminated. Personnel handling laundry must wear protective gloves. Laundry must be bagged at the location where it was used. If soaking through is likely, double bagging is required.
 - *Waste Management* - Employees are required to wear gloves when handling any infectious waste.
 - *Labeling* - A label showing the biohazard symbol will be affixed to all containers of infectious waste (i.e. biohazardous and medical waste), refrigerators, and freezers containing blood or other potentially infectious materials. The biohazard symbol must be black on an orange background.
 - *Transportation* - All specimens or containers of blood and tissue will be transported within a secondary container (e.g., plastic bag or other container having a liquid tight seal). These materials will be placed in a secondary container and labeled with the biohazard symbol prior to being taken into common areas.
 - *Food and Drink* - Eating, drinking, applying cosmetics or lip balm, and handling contact lenses are prohibited in laboratories and other work areas where blood or tissue, or other potentially infectious materials are present.

Please refer to the Appendix 7 – Safety Program Plan for more details.

6.0 Disposition of Data

All documents, data and study records collected for the purposes of this study will be stored indefinitely at the Fred Hutchinson Cancer Research Center. All researchers and staff with access to this information will follow procedures to prevent disclosure of information to anyone who is not an investigator on this study. This includes a pledge of confidentiality by all FHCRC, UW and PGS personnel; data handling procedures, network protection, password protection, proper storage and handling of all files and specimens, and secured facilities. A copy of the pledge of confidentiality is enclosed with this application.

6.1 Biostatistical Reviews

Biostatistical review of all project data will be conducted by the Core Co-Project Director, Garnet Anderson, PhD. Biostatistical reviews will be conducted to understand the behavior of new and known markers jointly. Analyses specific to data generated by the Core Laboratory and additional analyses of the consolidated project data will be performed. These analyses will include the following:

- Describing the joint and unique expression of p53, HER2/neu and Myc in tumor tissue, by disease status and stage.
- Describing the correlation between expression of p53, HER2/neu and Myc in peripheral blood and tumor tissue, by disease status and stage.
- Describing the relationship between serum CA-125 levels and the expression of p53, HER2/neu and Myc in tissue or peripheral blood.
- Describe the relationship between various clinical and epidemiological factors (e.g., disease stage, menopausal status, prior history of cancer, number of ovulatory cycles) and marker levels in blood.

7.0 Protocol Modification

Departure from protocol for individual subjects will not occur in this research study. The research investigators in this project acknowledge and accept their responsibility for protecting the rights and welfare of human research subjects and for complying with all human use and regulatory compliance as determined by the Institutional Review Office of the Fred Hutchinson Cancer Research Center and the Human Subjects Protection Division (HSPD). Research investigators will promptly report proposed changes in previously approved human subject research activities to both the IRO and HSPD. The proposed changes will not be initiated without IRB and HSPD review and approval.

7.1 Reporting of Serious and Unexpected Adverse Events

Serious and unexpected adverse experiences will be immediately reported by telephone to the USAMRMC Deputy Chief of Staff for Regulatory Compliance and Quality (301-619-2165) (non-duty hours call 301-619-2165 and send information by facsimile to 301-619-7803). A written report will follow the initial telephone call within 3 working days. Address the written report to the U.S. Army Medical Research and Materiel Command: ATTN: MCMR-RCQ, 504 Scott Street, Fort Detrick, Maryland 21702-5012.

8.0 Use of Information/Publications Arising From This Study

The personal identity of subject participants will not be revealed in any publication or release of results. All information/publications arising from this

study will be conducted in an ethical manner, as approved by the Institutional Review Office of the Fred Hutchinson Cancer Research Center.

9.0 Personnel to Conduct Project

Principal Investigator:	Nicole D. Urban, ScD	206-667-4677
Project Director	Garnet Anderson, PhD	206-667-4699
Project Director	Nancy Kiviat, MD	206-616-9740
Investigator	Charles Drescher, MD	206-587-0585
Investigator	Leona Holmberg, MD	206-667-6447
Investigator	Mary Anne Rossing, PhD	206-667-5041
Medical Monitor	Saul Rivkin, MD	206-386-2929

10.0 Signature of Principal Investigator

"I have read the foregoing protocol and agree to conduct the study as outlined herein."

Nicole Urban, ScD

Date

Appendix N

Specimen Transfer

- Specimen Delivery and Transfer Form

ORCHID – Specimen Tracking Form

Date: October 12, 2000

Specimens sent from: ORCHID Core Repository

- ☐ -20° C freezer
☐ -70° C freezer
☐ liquid nitrogen freezer

Specimen type Number

- ☐ Tissue _____
☐ Serum 10_____
☐ Other _____

Time of packaging: ____:____ AM / PM

Transported on:

- ☐ dry ice
☐ liquid nitrogen

Specimens packaged for delivery:

FOR NELSON LABORATORY – SPECIMENS FOR ORCHID PROJECT 2.

202774	100545
201601	202605
201775	100324
100724	100743
202746	100578

Specimens received:

Receiving laboratory: ☐ Kiviat laboratory

☐ Hellstrom laboratory (PNRI)

☐ M. Schummer laboratory (UW)

☐ Univ. of Washington: _____

☐ B. Nelson laboratory (VMMC)

☐ FHCRC: _____

☐ N. Disis laboratory (UW)

Time of delivery receipt: ____:____ AM / PM

Date of receipt: ____/____/____

☐ Contents above confirmed

☐ Contents different as noted: _____

Investigator/lab personnel signature: _____

TCS initials: _____

Appendix O

Characteristics of Participants

- Table 1: Characteristics of ORCHID Participants who completed the questionnaire
- Table 2: Clinical Characteristics by Outcome (N=299)
- Table 3: Clinical Characteristics by Outcome (N=244)
- Analysis of Project 1RT-PCR data
- Models 1-4

Table 1: Characteristics of ORCHID Participants who completed the questionnaire (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Demographics							
Age median (range)	52 (28-78)	58 (26-84)	45 (34-70)	63.5 (41-90)	65 (36-86)	57 (35-81)	55.5 (22-85)
Race N(%)							
Native American	2 (3.1)	1 (3.2)	1 (9.1)	0 (0.0)	1 (3.3)	0 (0.0)	1 (1.7)
Asian	3 (4.7)	2 (6.5)	1 (9.1)	4 (6.7)	1 (3.3)	2 (4.6)	2 (3.5)
Black	0 (0.0)	1 (3.2)	0 (0.0)	1 (1.7)	1 (3.3)	2 (4.6)	2 (3.5)
Caucasian	57 (89.1)	26 (83.9)	9 (81.8)	54 (90.0)	27 (90.0)	40 (90.9)	50 (86.2)
Other	2 (3.1)	1 (3.2)	0 (0.0)	1 (1.7)	0 (0.0)	0 (0.0)	3 (5.1)
Highest Educational Level N(%)							
<=11 th grade	1 (1.6)	1 (3.1)	2 (18.2)	5 (8.5)	2 (6.7)	1 (2.3)	0 (0.0)
High School Grad. or GED	7 (10.9)	7 (21.9)	2 (18.2)	14 (23.7)	8 (26.7)	8 (18.2)	21 (36.2)
Some college/votech	25 (39.1)	10 (31.3)	5 (45.6)	21 (35.6)	11 (36.7)	18 (40.9)	17 (29.3)
College graduate	16 (25.0)	3 (9.4)	2 (18.2)	12 (20.3)	6 (20.0)	11 (25.0)	10 (17.2)
Graduate school/advanced degree	15 (23.4)	11 (34.5)	0 (0.0)	7 (11.9)	3 (10.0)	6 (13.6)	10 (17.2)
Marital Status N(%)							
Currently Married	47 (74.6)	19 (59.4)	6 (54.6)	21 (35.0)	15 (50.0)	23 (53.5)	38 (65.5)
Other	16 (25.4)	13 (40.6)	5 (45.5)	39 (65.0)	15 (50.0)	20 (46.5)	20 (34.5)
Area of Birth N(%)							
USA	55 (88.7)	30 (93.8)	10 (90.9)	51 (86.4)	29 (96.7)	38 (90.5)	55 (94.8)
Europe	2 (3.2)	1 (3.1)	0 (0.0)	4 (6.8)	0 (0.0)	1 (2.4)	1 (1.7)
Asia	3 (4.8)	0 (0.0)	1 (9.1)	2 (3.4)	1 (3.3)	2 (4.8)	1 (1.7)
Other	2 (3.2)	1 (3.1)	0 (0.0)	2 (3.4)	0 (0.0)	1 (2.4)	1 (1.7)
Work Status N (%)							
Work full/part time	45 (71.4)	18 (56.3)	9 (81.8)	29 (48.3)	12 (40.0)	28 (63.6)	35 (60.3)
Retired	10 (15.9)	10 (31.3)	0 (0.0)	24 (40.0)	11 (36.7)	7 (15.9)	14 (24.1)
Other	8 (12.7)	4 (12.5)	2 (18.2)	7 (11.7)	7 (23.3)	9 (20.5)	9 (15.5)
Height Median (range)	64 (59-71)	65.5 (58-72)	67 (61-70)	64 (53-70)	65.5 (56-70)	65 (58-70)	65 (60-72)
Weight Median (range)	145 (98-280)	160 (112-250)	160 (140-200)	149.5 (100-275)	149 (105-260)	145 (95-350)	159.5 (104-270)
BMI Median (range)	24.3 (17.9-51.3)	25.8 (19.6-41.5)	24.9 (22.6-33.9)	25.4 (19.1-44.0)	25.8 (18.1-41.4)	25.7 (18.5-50.3)	26.3 (18.9-43.0)

Table 1(continued): Characteristics of ORCHID Participants (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Menstrual History							
Age at menarche median (range)	13 (9-17)	13 (9-17)	13 (12-17)	13 (11-16)	13 (11-15)	13 (9-16)	13 (9-17)
Ever Pregnant N(%)							
Yes	56 (87.5)	24 (77.4)	10 (90.9)	51 (86.4)	25 (83.3)	39 (90.7)	44 (75.9)
No	8 (12.5)	7 (22.6)	1 (9.1)	8 (13.6)	5 (16.7)	4 (9.3)	14 (24.1)
Age at First birth median (range)	24 (16-38)	23.5 (17-32)	24 (17-37)	22.5 (16-35)	24 (16-40)	23 (16-30)	23 (17-40)
Total # pregnancies > 6 months							
None	4 (7.4)	5 (20.8)	1 (10.0)	3 (6.0)	1 (4.4)	5 (14.3)	2 (4.7)
1	9 (16.7)	3 (12.5)	1 (10.0)	7 (14.0)	5 (21.7)	8 (22.9)	6 (14.0)
2	20 (37.0)	8 (33.3)	2 (20.0)	17 (34.0)	8 (34.8)	12 (34.3)	18 (41.9)
3	17 (31.5)	6 (25.0)	3 (30.0)	13 (26.0)	3 (13.0)	5 (14.3)	8 (18.6)
4+	4 (7.4)	2 (8.3)	3 (30.0)	10 (20.0)	6 (26.1)	5 (14.3)	9 (20.9)
Ever Breastfed N(%)							
Yes	33 (52.4)	9 (29.0)	6 (54.6)	29 (48.3)	12 (40.0)	19 (44.2)	23 (40.4)
No	30 (47.6)	22 (71.0)	5 (45.6)	31 (51.7)	18 (60.0)	24 (55.8)	34 (59.6)
Total months breastfed median (range)	8 (1-34)	7.5 (1-22)	8.5 (1-38)	6 (1-80)	7.5 (1-90)	7 (1-45)	7.5 (2-43)
Ever use birth control pills (BCP)							
Yes	47 (73.4)	20 (62.5)	7 (63.6)	32 (55.2)	19 (63.3)	28 (65.1)	36 (64.3)
No	17 (26.6)	12 (37.5)	4 (36.4)	26 (44.8)	11 (36.7)	15 (34.9)	20 (35.7)
Total months BCP median (range)	60 (0-324)	72 (0-276)	36 (12-144)	48 (0-300)	48 (0-216)	60 (0-336)	72 (12-240)
Hysterectomy N (%)							
Yes	18 (34.6)	16 (57.1)	3 (33.3)	29 (56.9)	12 (42.9)	20 (50.0)	20 (37.7)
No	34 (65.4)	12 (42.9)	6 (66.7)	22 (43.1)	16 (57.1)	20 (50.0)	33 (62.3)
Age at hysterectomy median (range)	50 (29-76)	43.5 (32-71)	63 (41-70)	52 (28-76)	53.5 (35-78)	48 (30-79)	49 (32-84)
Age at last period median (range)	48 (28-61)	44.5 (26-73)	44 (33-52)	48 (28-57)	50 (35-56)	46 (30-56)	47 (22-60)
Number ovulatory cycles median (range)	373 (78-578)	340 (55-727)	292 (124-428)	397 (13-557)	329 (199-543)	375 (91-510)	377 (51-575)
Ever used hormone replacement therapy N(%)							
Yes	31 (48.4)	20 (62.5)	4 (36.4)	41 (68.3)	12 (41.4)	23 (53.5)	29 (50.0)
No	33 (51.6)	12 (37.5)	7 (63.6)	19 (31.7)	17 (58.6)	20 (46.5)	29 (50.0)
Total months used HRT median (range)	48 (0-240)	84 (0-456)	6 (0-12)	96 (0-564)	42 (0-648)	96 (0-324)	36 (0-348)

Table 1(continued): Characteristics of ORCHID Participants (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/ Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Personal Medical History							
Ever diagnosed with diabetes N(%)							
Yes	3 (4.8)	5 (15.6)	2 (18.2)	1 (1.7)	4 (13.3)	3 (7.1)	3 (5.2)
No	59 (95.2)	27 (84.4)	9 (81.8)	57 (98.3)	26 (86.7)	39 (92.7)	55 (94.8)
Ever diagnosed with inflamed bowel synd. N(%)							
Yes	7 (11.1)	2 (6.9)	0 (0.0)	7 (12.1)	2 (6.7)	4 (10.3)	4 (7.0)
No	56 (88.9)	27 (93.1)	11 (100.0)	51 (87.9)	28 (93.3)	35 (89.7)	53 (93.0)
Ever diagnosed with fibroids in uterus N(%)							
Yes	34 (54.8)	22 (68.8)	4 (36.4)	19 (33.3)	6 (20.7)	13 (31.7)	23 (40.4)
No	28 (45.2)	10 (21.3)	7 (63.6)	38 (66.7)	23 (79.3)	28 (68.3)	34 (59.7)
Ever diagnosed with endometriosis N(%)							
Yes	6 (10.9)	3 (9.7)	0 (0.0)	5 (9.8)	4 (14.3)	4 (11.1)	8 (14.3)
No	49 (89.1)	28 (90.3)	9 (100.0)	46 (90.2)	24 (85.7)	32 (88.9)	48 (85.7)
Ever diagnosed with benign breast disease N(%)							
Yes	13 (21.7)	14 (46.7)	3 (27.3)	10 (18.5)	7 (23.3)	15 (36.6)	10 (17.2)
No	47 (78.3)	16 (53.3)	8 (72.3)	44 (81.5)	23 (76.7)	26 (63.4)	48 (82.7)
Ever diagnosed with polycystic ovarian dis. N(%)							
Yes	3 (6.3)	0 (0.0)	2 (25.0)	2 (5.0)	1 (3.6)	2 (5.6)	3 (6.1)
No	45 (93.8)	23 (100.0)	6 (75.0)	38 (95.0)	27 (96.4)	34 (94.4)	46 (93.9)
Ever diagnosed with ovarian cyst N(%)							
Yes	27 (51.9)	16 (64.0)	6 (85.7)	10 (23.3)	3 (10.7)	19 (50.0)	20 (38.5)
No	25 (48.1)	9 (36.0)	1 (14.3)	33 (76.7)	25 (89.3)	19 (50.0)	32 (61.5)
Ever diagnosed with breast cancer N(%)							
Yes	8 (12.5)	3 (9.4)	0 (0.0)	5 (8.5)	0 (0.0)	5 (11.4)	2 (3.5)
Age diagnosed with breast cancer median (range)	48.5 (32-59)	50 (35-53)	0	51 (36-62)	0	54 (53-63)	46 (30-62)

Table 1(continued): Characteristics of ORCHID Participants (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/ Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Family History: # 1 st degree relatives with:							
Breast Cancer N(%)							
0	52 (81.3)	24 (75.0)	10 (90.9)	52 (86.7)	25 (83.3)	39 (88.6)	52 (89.7)
1+	12 (18.8)	8 (25.0)	1 (9.1)	8 (13.3)	5 (16.7)	5 (11.4)	6 (10.3)
Ovarian Cancer N(%)							
0	63 (98.4)	32 (100.0)	11 (100.0)	58 (96.7)	29 (96.7)	42 (95.5)	55 (94.8)
1+	1 (1.6)	0 (0.0)	0 (0.0)	2 (3.3)	1 (3.3)	2 (4.6)	3 (5.2)
Prostate Cancer N(%)							
0	59 (92.2)	28 (87.5)	10 (90.9)	54 (90.0)	27 (90.0)	37 (84.1)	55 (94.8)
1+	5 (7.8)	4 (12.5)	1 (9.1)	6 (10.0)	3 (10.0)	7 (15.9)	3 (5.2)
Lifestyle							
Smoking N(%)							
Current smoker	6 (9.4)	3 (9.7)	1 (9.1)	8 (13.3)	1 (3.5)	8 (18.6)	3 (5.2)
Former smoker	21 (32.8)	11 (35.5)	4 (36.4)	19 (31.7)	14 (48.3)	12 (27.9)	22 (37.9)
Never smoked	37 (57.8)	17 (54.8)	6 (54.6)	33 (55.0)	14 (48.3)	23 (53.5)	33 (56.9)
Ever drink alcohol N (%)							
Yes	47 (73.4)	24 (75.0)	8 (72.3)	38 (63.3)	21 (72.4)	26 (60.5)	35 (61.4)
No	17 (26.6)	8 (25.0)	3 (27.3)	22 (36.7)	8 (27.6)	17 (39.5)	22 (38.6)

Table 2: Clinical Characteristics by Outcome (N=299)

Characteristic	Normal	Benign Ovarian Pathology	LMP/Borderline Ovarian Tumor	Outcome Ovarian Cancer	Other Cancer	Unknown	No Tissue Donation
Presenting symptoms N (%)							
Pain	26 (40.6)	10 (31.3)	2 (18.2)	43 (71.7)	6 (20.0)	20 (45.6)	2 (3.5)
Distention	12 (18.8)	6 (18.8)	3 (27.3)	39 (65.0)	5 (16.7)	20 (45.6)	2 (3.5)
Bleeding	23 (35.9)	6 (18.8)	3 (27.3)	4 (6.7)	17 (56.7)	5 (11.4)	0 (0.0)
Fatigue	1 (1.6)	2 (6.3)	0 (0.0)	10 (16.7)	2 (6.7)	5 (11.4)	0 (0.0)
Dyspepsia	2 (3.2)	0 (0.0)	2 (18.2)	14 (23.3)	2 (6.7)	9 (20.5)	0 (0.0)
Weight Change	5 (7.1)	2 (6.3)	1 (9.1)	13 (21.7)	2 (6.7)	5 (11.4)	1 (1.7)
Bladder Changes	13 (20.3)	3 (9.4)	0 (0.0)	6 (10.0)	1 (3.3)	7 (15.9)	0 (0.0)
Bowel Changes	7 (10.9)	3 (9.4)	2 (18.2)	13 (21.7)	1 (3.3)	8 (18.2)	1 (1.7)
Other	11 (17.2)	2 (6.3)	1 (9.1)	10 (16.7)	2 (6.7)	6 (13.6)	1 (1.7)
Pre-operative CA-125 median (range)	12 (2-84)	33.6 (9.6-58)	109.9 (4-1587)	427 (8.8-9060)	338 (338-338)	44.5 (5.3-4667)	611.2 (578.3-644)

Table 3: Clinical Characteristics by Outcome (N=244)

Characteristic	FIGO Stage			
	0	I	II	IV
Grade N(%)				
Well differentiated	1 (25.0)	3 (42.9)	0 (0.0)	0 (0.0)
Moderately differentiated	1 (25.0)	4 (57.1)	0 (0.0)	2 (28.6)
Poorly differentiated	2 (50.0)	0 (0.0)	1 (100.0)	5 (71.4)
Histology				
Ovarian Cancer	2 (1.2)	8 (47.1)	2 (50.0)	7 (70.0)
LMP	0 (0.0)	9 (52.9)	2 (50.0)	0 (0.0)
Unknown	37 (22.3)	0 (0.0)	0 (0.0)	2 (16.7)
HER-2 median (range)	0 (0-8.2)	0 (0-3.3)	1.1 (0.6-2.4)	0.1 (0-2.7)

Analysis of Project 1RT-PCR data

Gene	NORMAL			CANCER			Modified		Ranking	
	Mean	StdDev	Minimum Maximum	Mean	StdDev	Minimum Maximum	T	T	Original	Revised
SLPI	0.00	0.00	0.00	58.05	79.41	0.00	312.32	2.83 Infinity	15	1
HE4	0.04	0.04	0.01	32.04	47.13	0.03	190.62	2.63 2282.48	20	2
Mesothelin	1.01	0.89	0.01	477.99	1125.67	0.00	4453.90	1.64 1519.46	43	3
Mucin1	0.11	0.12	0.02	31.44	50.97	0.34	206.90	2.38 719.25	26	4
Folate BP	1.06	0.99	0.10	79.56	86.19	1.71	271.80	3.53 223.45	7	5
CD24	0.75	1.62	0.00	109.62	172.49	0.90	690.26	2.44 189.22	25	6
Keratin8	0.20	0.14	0.02	7.35	11.06	0.08	40.34	2.50 144.47	24	7
ESE-1	0.02	0.05	0.00	1.73	1.91	0.00	5.56	3.47 101.67	9	8
Ku80	0.12	0.48	0.00	11.79	31.22	0.00	110.14	1.45 68.41	49	9
Lipocalin2	0.50	1.09	0.05	24.27	36.34	1.38	148.34	2.53 61.41	23	10
BRCA1	0.13	0.10	0.00	2.29	6.01	0.06	23.76	1.39 58.95	50	11
oviductGP	0.11	0.17	0.01	3.02	10.35	0.00	40.39	1.09 48.96	57	12
p53	0.41	0.51	0.00	6.87	4.60	0.00	10.00	5.41 35.92	1	13
Ferritin H	4.20	5.85	1.48	46.05	158.03	0.32	617.05	1.03 20.20	60	14
Enolase	7.03	4.45	1.31	37.95	24.07	1.51	92.42	4.90 19.61	3	15
KIAA0762	5.12	2.72	1.47	22.58	20.08	0.23	56.21	3.34 18.14	10	16
T000M-07-F19	2.33	1.15	0.39	9.61	10.30	0.00	39.72	2.72 17.92	19	17
Ryudocan (no T069c)	4.21	3.21	0.79	22.22	21.76	0.00	70.50	3.18 15.84	12	18
RIG-E	97.32	45.01	16.59	348.61	244.00	16.47	854.06	3.93 15.76	5	19
p27	23.93	31.65	0.92	198.95	140.58	17.94	448.55	4.72 15.61	4	20
TGF beta 1	0.10	0.08	0.02	0.55	1.63	0.03	6.43	1.07 15.55	59	21
T000-09-c15	1.87	1.49	0.12	8.81	12.63	0.21	50.63	2.12 13.15	32	22
MR	0.75	0.72	0.00	3.54	1.94	0.00	7.06	5.24 10.88	2	23
GAPDH	14.16	8.35	0.68	43.41	27.92	1.76	94.29	3.91 9.89	6	24
IGF BP2	10.99	7.72	0.25	36.64	45.76	0.00	144.46	2.14 9.38	31	25
MCAF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.64 9.32	44	26
BRCA2	1.07	0.68	0.00	3.31	2.65	0.30	8.79	3.18 9.26	11	27
GPR39	0.01	0.02	0.00	0.05	0.07	0.00	0.25	2.80 8.82	16	28
PAX2	0.00	0.01	0.00	0.04	0.07	0.00	0.21	1.89 8.04	34	29
T000M-171-111	2.03	0.93	0.83	4.60	3.81	1.04	16.52	2.54 7.76	22	30
MAT-1	0.99	0.54	0.20	2.41	1.84	0.00	6.49	2.90 7.38	14	31
T000M-97-E17	0.39	0.23	0.04	0.97	0.96	0.10	3.36	2.29 7.18	27	32

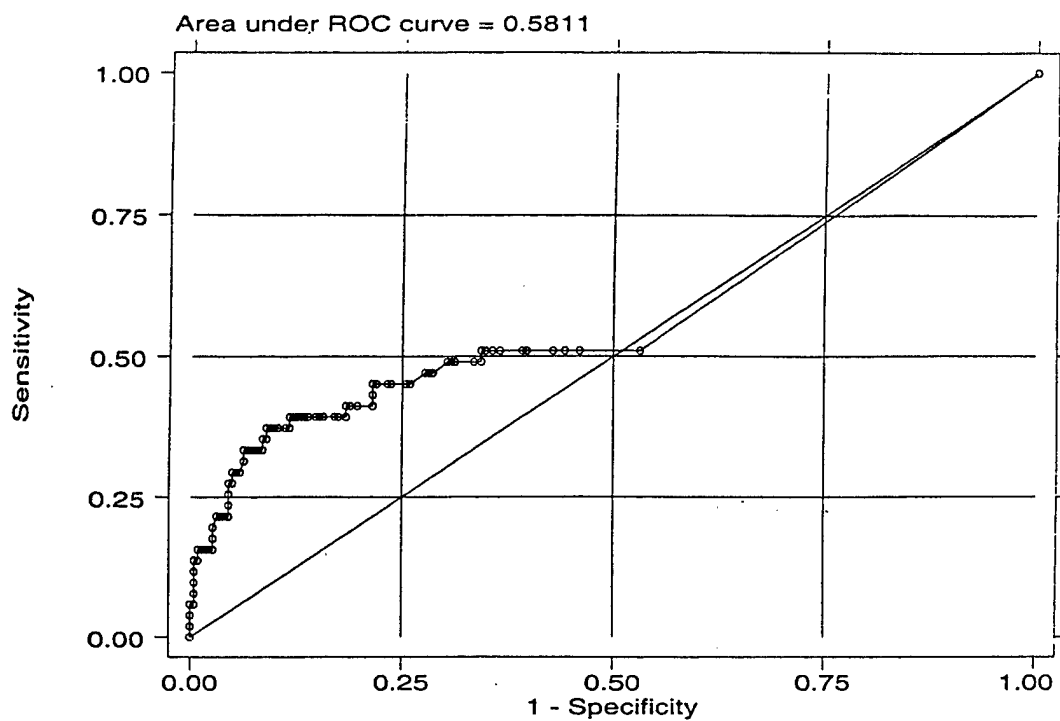
Analysis of Project 1RT-PCR data

Gene	NORMAL				CANCER				Modified		Ranking	
	Mean	StdDev	Minimum	Maximum	Mean	StdDev	Minimum	Maximum	T	T	Original	Revised
T000M-06-B19	1.71	1.60	0.05	6.67	5.58	6.65	0.18	24.15	2.20	6.81	29	33
T000M-60-F20	0.03	0.05	0.00	0.14	0.14	0.14	0.00	0.48	2.76	6.15	17	34
H2N (2nd ref)	2.10	1.92	0.00	5.81	6.21	8.55	0.00	31.37	1.82	6.02	36	35
T000M-106-H01	0.34	0.43	0.06	1.97	1.23	0.89	0.15	2.82	3.51	5.81	8	36
ProgBP	0.53	0.36	0.05	1.70	1.24	1.21	0.38	4.26	2.21	5.55	28	37
IGF2	0.74	0.68	0.00	3.09	2.03	3.21	0.00	10.55	1.53	5.36	47	38
14.3.3	3.48	3.45	0.31	12.74	9.78	10.82	0.00	42.54	2.16	5.16	30	39
hose-060c2403	8.83	3.20	1.90	14.34	14.63	8.17	1.77	37.39	2.58	5.11	21	40
T000M-76-K21	11.19	10.25	0.14	40.41	29.54	20.57	3.43	66.00	3.13	5.05	13	41
MAGE E1	69.86	38.59	14.65	120.06	129.90	119.34	2.51	451.87	1.86	4.39	35	42
1820-02-o1102	0.69	0.38	0.19	1.55	1.26	1.21	0.00	4.29	1.76	4.22	39	43
ZFP161	0.28	0.21	0.00	0.75	0.58	0.63	0.00	2.15	1.80	4.17	37	44
T000M-139-A09	0.79	0.25	0.34	1.19	1.16	0.46	0.26	1.82	2.72	4.13	18	45
AA447275	0.34	0.17	0.13	0.72	0.57	0.67	0.09	2.48	1.29	3.81	54	46
KIAA0991	4.30	4.52	0.00	19.43	10.33	11.02	0.26	44.63	1.98	3.76	33	47
KIAA0512	2.03	1.20	0.35	4.38	3.52	3.19	0.15	12.02	1.70	3.48	40	48
Calgizarin	13.25	9.39	2.37	35.07	24.67	25.34	0.00	88.03	1.65	3.43	42	49
KIAA0952	1.56	1.11	0.00	3.59	2.79	2.45	0.31	7.40	1.79	3.13	38	50
T000M-187-K19	19.32	75.38	0.00	311.76	101.88	339.95	0.00	1316.90	0.92	3.09	63	51
1-4	2.24	4.21	0.14	15.55	6.76	18.78	0.10	73.67	0.91	3.03	64	52
T000M-26-I14	0.36	0.23	0.00	0.88	0.58	0.48	0.00	1.83	1.67	2.80	41	53
H53727	0.18	0.28	0.00	1.10	0.46	0.63	0.00	2.07	1.53	2.72	46	54
T000M-14-M15	0.78	0.47	0.03	1.62	1.18	1.25	0.18	4.36	1.18	2.41	56	55
PTEN	5.52	5.24	0.00	17.34	9.96	12.85	0.00	48.92	1.25	2.39	55	56
MDC15	0.51	0.75	0.00	2.54	1.11	1.27	0.01	3.70	1.60	2.25	45	57
E16	14.14	35.87	0.00	149.79	40.92	96.40	0.00	380.96	1.02	2.11	61	58
T000M-34-L01	12.10	0.95	9.96	13.65	12.77	1.46	10.47	15.45	1.51	1.99	48	59
S31iii125-2	18.23	13.12	1.21	49.42	27.04	28.84	3.56	120.62	1.09	1.89	58	60
98118-D01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	1.36	1.85	51	61
KIAA0263	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.03	0.59	1.79	69	62
H2N	0.41	0.51	0.00	1.00	0.73	1.75	0.00	5.00	0.69	1.79	68	63
Actin	35.87	40.18	0.25	126.74	59.56	56.99	0.10	140.84	1.34	1.66	52	64

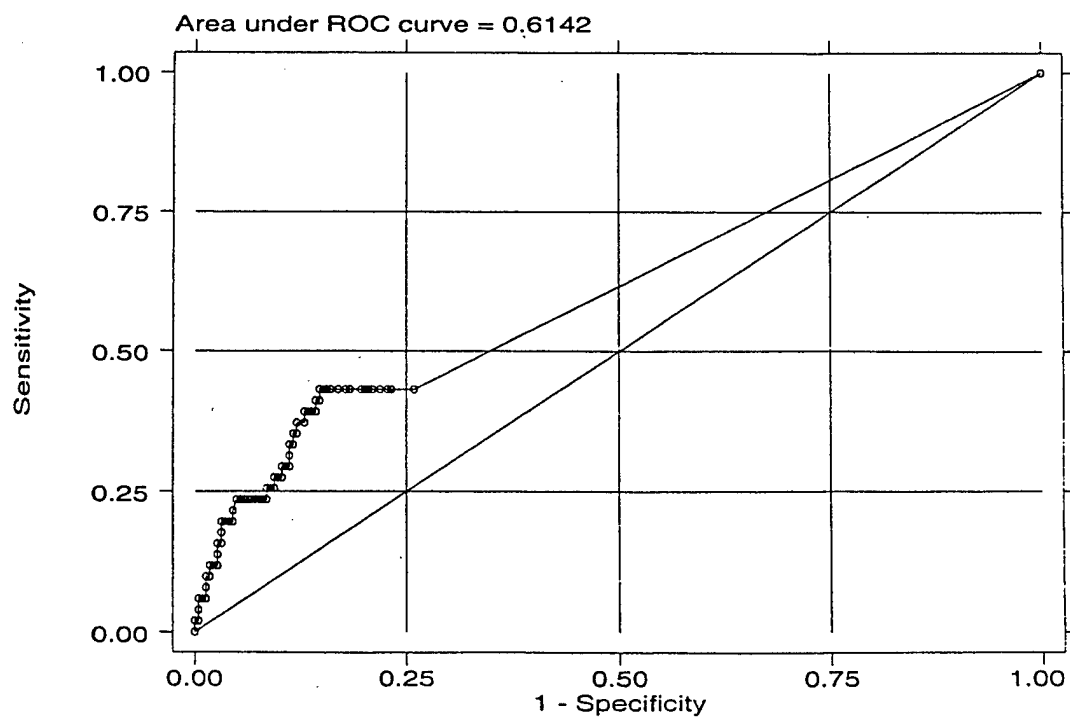
Analysis of Project 1RT-PCR data

Gene	NORMAL				CANCER				Modified		Ranking	
	Mean	StdDev	Minimum	Maximum	Mean	StdDev	Minimum	Maximum	T	T	Original	Revised
S31iii125	8.01	6.75	0.60	26.12	11.07	6.57	1.77	25.80	1.30	1.28	53	65
PLTP	123.14	75.57	12.45	312.87	156.25	147.00	5.01	462.15	0.79	1.24	67	66
T000M-16-A21	1.30	0.99	0.10	4.41	1.73	1.34	0.11	5.57	1.01	1.21	62	67
Bamacan	3.40	2.53	1.22	11.47	4.32	3.76	0.50	14.99	0.80	1.03	66	68
TRC8	4.13	3.82	0.65	16.63	5.22	3.13	0.85	10.25	0.89	0.81	65	69
SP5 (noT069c)	20.42	6.73	6.93	30.78	22.18	12.94	0.00	57.31	0.47	0.74	71	70
c-myc	1.47	2.84	0.00	11.43	2.06	3.05	0.00	10.78	0.57	0.59	70	71
T000M-31-N08	2.16	1.91	0.00	6.26	2.34	2.05	0.00	7.64	0.26	0.27	72	72
N000-11-E24	0.02	0.02	0.00	0.08	0.02	0.04	0.00	0.15	0.10	0.15	73	73
Kadereit	0.73	0.57	0.13	2.26	0.65	0.56	0.06	2.05	-0.41	-0.41	74	74
SAS	5.39	16.34	0.00	65.38	1.90	2.29	0.05	7.70	-0.87	-0.60	76	75
BA46	55.59	31.06	7.09	100.98	48.50	50.31	7.56	187.37	-0.47	-0.64	75	76
TLE4	2.24	1.97	0.19	6.84	0.89	0.97	0.04	3.66	-2.49	-1.93	77	77
fij10561 noStandard	287.76	167.22	0.00	578.04	139.81	155.09	0.00	447.09	-2.60	-2.50	78	78

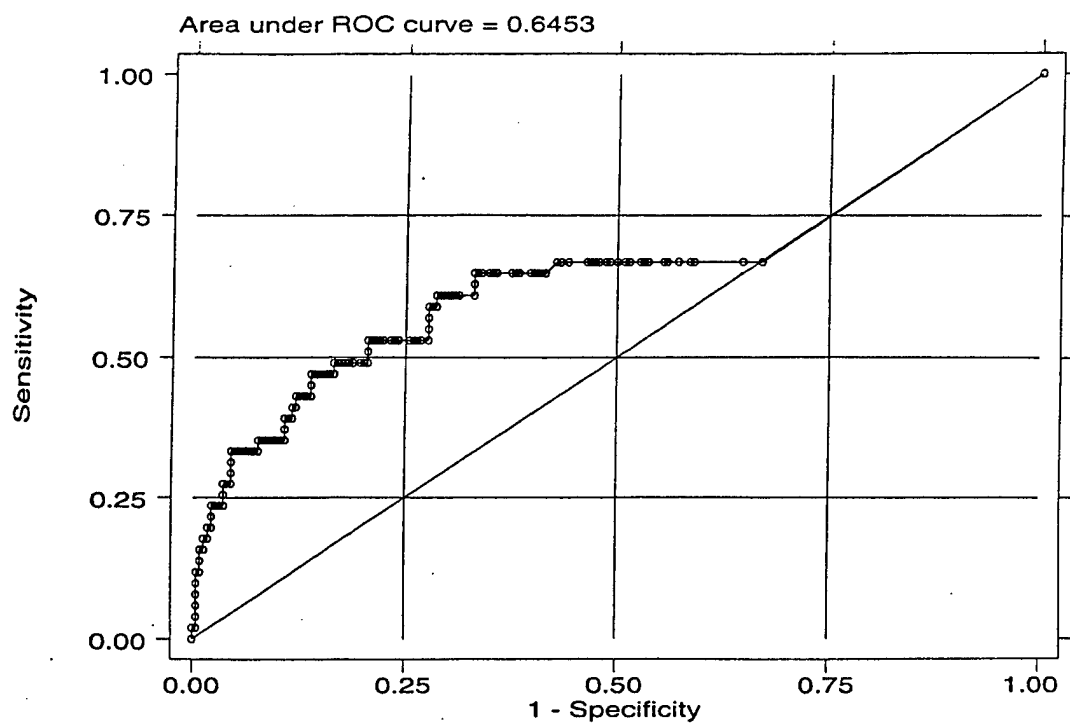
Model 1: $\text{logit}(p) = \beta_0 + \beta_1 \ln(p53 + 1)$



Model 2: $\text{logit}(p) = \beta_0 + \beta_1 \ln(h2n + 1)$

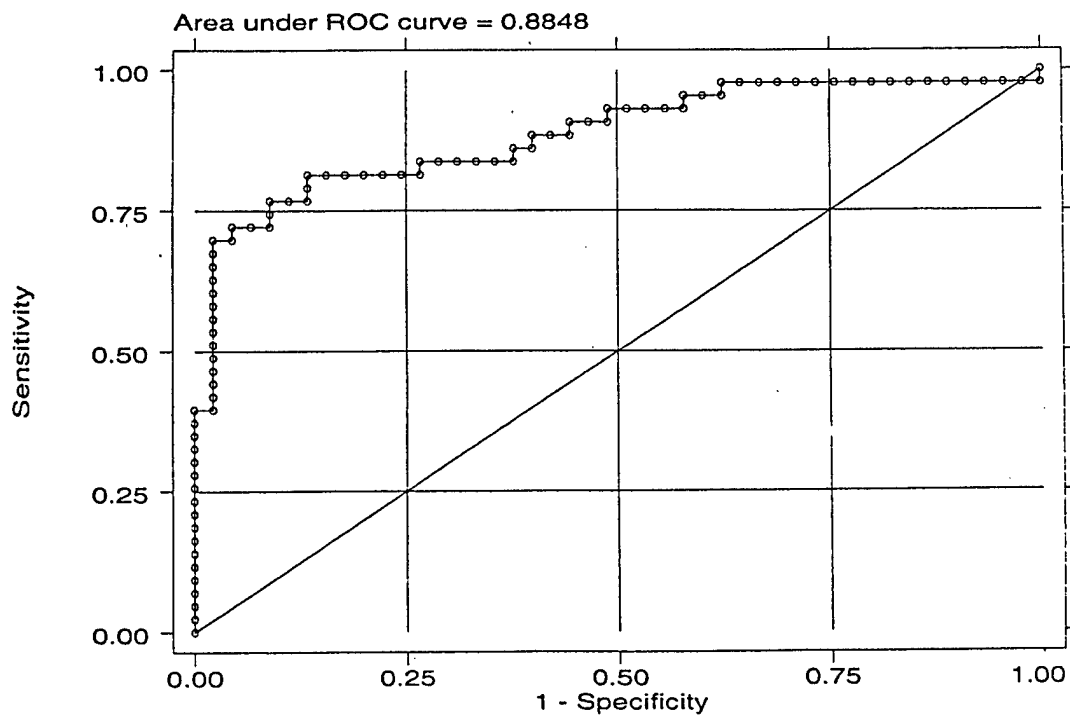


Model 3: $\text{logit}(p) = \beta_0 + \beta_1 \ln(p53 + 1) + \beta_2 \ln(H2N + 1)$



5 " 2 1 2 "

$\text{logit}(p) = \beta_0 + \beta_1 \text{Age} + \beta_2 \ln(\text{CA-125} + 1)$



Model 4: $\text{logit}(p) = \beta_0 + \beta_1 \text{Age} + \beta_2 \ln(\text{CA-125} + 1) + \beta_3 \ln(\text{p53} + 1) + \beta_4 \ln(\text{H2N} + 1)$

